STATISTICS
HOW TO HANDLE THEM

First Indian Edition

STATISTICS

HOW TO HANDLE THEM

A Practical Handbook for Students and Businessmen with a Special Chapter on BUSINESS FORECASTING

BY

A. K. SUR, M. A.,
Editor, *Calcutta Stock Exchange Official Year Book*

D. B. TARAPOREVALA SONS & Co.
210, Hornby Road, Fort - - - BOMBAY

Printed by S. Ramu at The Commercial Printing Press, (Proprietors : Tata Sons, Ltd.), 105, Cowasji Patel Street, Fort, Bombay and Published by Jal Hirji Taraporevala for D. B. Taraporevala Sons & Co., 210, Hornby Road, Bombay.

PREFACE

These are days of planning, and in view of the fact that statistics play an important role in national organization and planning, a book like this has at this moment a special interest.

The author has spent the past ten years of his life in the Stock Exchange doing statistical work of an intensive nature. During this period he has always felt the need of presenting to the businessmen in general an exposition of statistical method more rationalized than usually found in textbooks on the subject. The text, therefore, develops in the simplest possible manner, the statistical techniques, first for the treatment of numerical data and then for the drawing of inferences therefrom. In the development of this treatise the author has had always the needs and requirements of the businessman in view. The author realizes that the businessman may not have the required preparation in algebra and calculus necessary for various statistical calculations. On this assumption the author has all through the work given calculations involving the use of arithmetic only.

At every stage the methods are illustrated in their application to a variety of economic and business problems. Illustrative materials and problems have been chiefly taken from the field of Indian economics and business. In certain cases, however, assumed or hypothetical data have been used, and where this is done it is indicated by a footnote.

To initiate the businessman in calculation by logarithms an explanatory chapter giving the practical aspects of it has been given at the end of the book.

The author is thankful to his numerous friends and associates in Clive Street for stimulating him with encouragement all through the work.

A. K. SUR.

CONTENTS

CHAPTER I

STATISTICS IN BUSINESS

When used in the plural, 'Statistics' means numerical facts or data in any department of enquiry placed in relation to each other. In its singular sense, however, it refers to the subject itself, which is defined in the *Oxford Dictionary* as "the department of study that has for its object the collection and arrangement of numerical facts or data, whether relating to human affairs or to natural phenomena." This is however, a narrow definition of the Science of Statistics. For, Statistics concerns itself not merely with the collection and arrangement of numerical facts or data, but with their analysis and interpretation as well.

Taken in isolation, statistical data fail to tell their own stories. But when properly collated and co-ordinated for comparison with appropriate or cognate groups of items they become voluble and shed much light on problems that otherwise would remain obscure and unintelligible. In this way statistical data have made intelligible to us problems in many a practical affair of life, that otherwise would have remained devoid of any significance. Indeed, "the use of statistics in the business of running the country

through its political, commercial and social institutions —in those activities that determine the health, wealth and happiness of mankind— is the oldest and the most considerable use."

In the business sphere particularly statistics have their uses in diverse ways. Business itself, it may be noted, is founded upon estimates and probabilities. The businessman sizes up his production on the basis of probable demand for same. Such estimates are always based upon past records and experience, as also upon the changing tastes of the times, and if there be any error or blunder in the estimate, the businessman is apt to come to grief. Success in business rests upon estimates approximating as nearly as possible to actual results.

Discrepancy between estimates and results cannot be eliminated unless the estimates are based upon some scientific methods. And it is the concern of the Science of Statistics to point out the correct methods.

It is for this reason that we find progressive businessmen the world over realizing more and more the importance of setting up statistical branches as necessary adjuncts to their business organizations.

It is the function of such statistical research departments to collect, collate and co-ordinate facts and figures, not only relating to the particular business of the house, but in regard to competitive businesses too and of trade and finance generally. Data thus collected help the business houses concerned in exploring new markets and fresh avenues of income, as also in eliminating competition. For the internal organization of the house itself such data are of great value for the solution of many a problem of production, selling, management and budgetary control.

Furthermore, such statistical investigations provide the firm concerned with a life and death test of its own progress, for the nature and effects of any changes observed, when properly analyzed and checked, are helpful in keeping the machine moving much more smoothly or saving it from peril.

CHAPTER II

GATHERING OF DATA

The first step in all statistical investigations is, of course, the gathering of appropriate data for purposes of estimation. Utmost care should be devoted to this aspect of the work, in as much as if the data are imperfect or not judiciously selected, the results or conclusions are likely to be erroneous and misleading. As a point of fact, data to be of real value for statistical purposes, must be *precise*, *accurate* and *stable*. To ensure this, terms to which the figures relate should have a precise connotation, and that connotation or definition should not be changed or modified at any stage during the whole process of gathering the data. For instance, although such terms as 'manufacturing concerns,' 'trade,' 'profit,' 'working expenses,' etc., have different significance in common usage, yet when used for statistical treatment, they should have a definite connotation, and that connotation should be strictly adhered to all through the same enquiry. In other words, variables (*i.e.*, entities like 'profits,' 'sales,' 'outputs,' etc., which assume different values in different periods and circumstances)

should remain static in significance and connotation all through the enquiry. Lastly, we must be definitely precise as to what particular region or place the statistics refer, and to what instant or period of time they relate.

Statistical data can be gathered either from *published sources* of information or by *ad hoc enquiries*. Published sources abound in many fields of economic investigations. Firstly, there are the official blue books periodically issued by the various government departments, generally as a by-product of certain administrative operations. Secondly, the reports of the various Royal Commissions and similar other statutorily appointed bodies embody the results of many statistical enquiries specifically undertaken by them. Thirdly, there are the reports of the Reserve Bank of India and of various non-official organizations like the Chambers of Commerce, the Stock Exchanges, the Central Jute Committee, the Indian Jute Mills Association, the Millowners' Associations, and similar other bodies. Trade gazettes and some of the financial papers also feature many statistical Index Numbers specially prepared by them.

As regards *ad hoc* enquiries or special types of investigations made by individuals, the

Questionnaire method appears to be best. 'Questionnaire forms' for such purposes must be compiled with utmost care and thought. These should be as simple as possible, and the questions should be so carefully framed as to elicit either a precise answer like "Yes" or "No," or a definite number. Where possible ambiguity may arise all questions should be clearly defined. Further, the questions asked must be of an inoffensive character, and must not be so framed as to deter the person to whom they are addressed from answering all the facts about them candidly. Lastly, the form should contain enough blank spaces to accommodate the answers even when penned in large handwriting. All such 'questionnaire forms' should be headed with appropriate instructions for filling in of the data, to make mistakes difficult.

Classes of Data

For all practical purposes, statistical data may be classified under two heads: (*i*) a *Time Series*, that is to say, a series in which items or observations are distributed in relation to some unit of time. This includes such items as monthly totals of jute mills production, company profits or crop production year by year, weekly earnings of railways and so forth.

Time series again are of two kinds. In the first kind the figures relate to some quantity which is measured at a particular instant of time, *e. g.*, the census of population figures in a country on a given date in a particular year. And in the second series the figures relate to some particular quantity in a number of time intervals, *e. g.*, the figures of total production of jute manufactures month by month. (*ii*) *A Frequency Distribution*, that is to say, a series in which the items of observations are distributed with respect to some physical characteristics, *e. g.*, distribution of jute mills according to the number of looms, the distribution of companies according to the size of capital or profits, the distribution of wage earners according to weekly wages received, and so forth. Data of a frequency distribution may be either of (*a*) the *Continuous* type *i. e.* one that may have any number of values ranging between the lowest and the highest, such as farm costs of production; or of the (*b*) the *Discrete* type *i. e.* one with a value distinct and separate, as, for example, the number of workers in a factory.

Census & Sample Enquiries

When the whole information about the subject of enquiry is collected we call it a *Census*

enquiry, but when only a part of it is sought we call it a *Sample* enquiry. In the study of business or economic problems, sample enquiry is generally found to be more expedient than census enquiry, because it saves time, money and labour. The process involved in selection of data for a sample enquiry is known as *Sampling.* Such a sample is generally drawn purely at random from the field of enquiry, but the one thing that is to be always clearly borne in mind when a sample is drawn, is that the examples picked are really representative of the entire field of investigation. A method that is usually employed for the random sampling process is either the lottery system or the system of selecting every tenth example (for instance, from a directory), or pricking the required number of cases in a list while blindfolded. One advantage of this method over that of personal selection is that it does not lead to unconscious bias. The maxim to be followed in this connection is that larger the number of representative cases or examples, the more accurate and satisfactory will be the standard.

A question that may naturally spring up in this connection is : what warrant is there that

estimates made from a ' sample ' are as good as that from the ' total population ' (the entire field of enquiry is in Statistics known as the ' Population ' or the ' Universe ')? Mathematicians have proved that " if a moderately large number of items be chosen at random from among a large mass, such numbers are on the average almost sure to have the same characteristics of the large group, and the data so obtained can be safely used as a base for comparison with all other examples of the same kind." This theory is known as the *Theory of Probability* or the *Law of Statistical Regularity.* (In such selection of cases, it should however be seen that the examples picked are really representative of the entire field of investigation).

Now, in course of our enquiry we may meet with some items or observations that would manifest unusual or abnormal values. But by another law known as the Law of the *Permanence of Small Numbers* we know that " when a particular or unusual characteristic occurs in a properly selected sample it may be expected that this same characteristic is likely to be present in the entire group from which the sample was taken in the same proportion

as it is present in the sample itself. In other words, the characteristic is permanent, being present in every group similarly selected and composed of the same number of individuals as the original sample."

Again, according to the *Law of Inertia of Large Numbers*, variations in one direction are apt to be offset or equalized by variations in another direction. The Moving Averages (see Ch. V) are particularly useful where such abnormalities are conspicuously evident. Lastly, it should be noted that there is a distinct relationship between the size and degree of precision of a sample. As a point of fact, the degree of precision of a sample increases as the number of items in the sample is increased, and that it varies in accordance with the square root of the number of items constituting the sample.

CHAPTER III

ARRANGING THE DATA

Statistical data, in their crude form, having been gathered, the statistician's next job is to set them out in tabular form in such a manner as to make them suitable for appropriate comparison and bring out their significant features before the reader's eye. When tabulating the data the problem for which the statistical enquiry is being conducted should be uppermost in the statistician's mind, and with that specific problem in mind he should devise how best the data can be presented to make them clear, convenient, intelligible and readable. In brief, the effectiveness of presentation and the immediacy of need should be the cardinal principles of all tabulation work. (Be it noted, however, that after the data for the immediate need have been tabulated, the original material should not be thrown away as the same may serve a different useful purpose in future).

Rules for Tabulation

(1) The statistical table should not be made too large and cumbersome, or of a size that is difficult for the eye to catch its significance at a glance. When however a large table becomes inevitable, it should be split up into sections,

and the different sections summed up in a separate subsidiary table. In this connection, Dr. Rhodes in his *Elements of Statistics* observes that a golden rule is to make out two tables instead of one if there is the least fear that one would contain such a mass of material as would tend to make the essential facts therein obscure and hidden.

2. The number of sections into which a column should be split up would, of course, depend upon the discretion of the statistician, but attention should always be paid to the relative importance of the various data, and the principle to be followed in this respect is to make the less important follow the more important. The usual method that is followed is to arrange the sub-divisions on the basis of some well-defined principles (*e. g.*, period, size, merit, function, alphabetical, geographical or spatial, species, etc.) in vertical form in the first column, and put the respective units of measurements against them horizontally.

3. All comparative figures should be placed in a vertical row, for a horizontal line of figures is more tiresome to the eye. Where the totals are the things that principally concern us, they should be put at the head of the vertical formation, or on the left side of a horizontal

tabulation, if of course, the comparables have been put horizontally.

4. Percentages loom large in business statistics and they are oftentimes incorporated into tabular statements. Where percentages are used due care must be given to their calculation. They should be correctly worked out to a definite place of decimals, and where the same base is used they can be checked by adding the percentages, since they would collectly amount to 100 per cent (though in actual practice a minor fractional variation is noticeable, but that is negligible). When however percentages are worked on different bases, they should not be added as above. In this case, the percentage of the total should be calculated on the actual totals of the columns concerned, as is shown in the table below :

Table I.—Gross Earnings & Working Expenses of the McLeod Group of Railways for the Year Ended 31-3-45.

Railways	Working Expenses	Gross Earnings	Per cent of Expenses to Gross Earnings
	Rs.	Rs.	
Ahmadpur-Kutwa ...	1,33,817	1,23,187	108·63
Bankura Damodar ...	1,69,162	1,49,785	112·93
Burdwan Kutwa ...	1,35,330	1,63,976	82·53
Kalighat Falta ...	2,20,773	2,14,277	103·03
Katakhal Lalabazar ...	34,282	68,564	50.00
Total...	6,93,364	7,19,789	96·33

When both actual figures and percentages or averages are placed in the same table, they should appear in close juxtaposition to each other.

5. If there be any gap against any item in any of the columns (that is to say, where such figures are not available), the totals thereof are not to be made up, or are to be shown in a different type from the rest.

6. Where numbers are composed of many digits, the terminal digits may be dropped by use of approximation*, but it should be noted that the figures do not lose their importance and value thereby. For instance, where the figures may be used for further tabulation or enquiry, they should be shown always in their entirety in actual numbers.

7. Large numbers should be separated by commas, and numbers in which fractional parts are involved, should be clearly separated by decimal points. Again, all such commas and decimal points in a column must be kept in proper alignment.

*Such methods of approximation enable us to obtain correct results in regard to the calculation of percentages to the nearest 1 per cent.

8. Finally, all statistical tables must be complete and self-explanatory. They should, as a point of fact, be accompanied by such explanatory notes as would leave no possible ambiguity in the interpretation of the meanings of the tables. They must have proper labels or titles to them at the top, and footnotes should be used to show sources of data and to point out factors of importance that cannot be otherwise incorporated in the body of the table.

Frequency Tables

When the data belong to a Frequency Distribution, there will arise the necessity of arranging the data according to their respective magnitudes, into what is known as a *Frequency Table*. Thus, if there be a group of 419 industrial undertakings, and of these 20 earn a profit of not over Rs. 5,000 per annum; 30 earn a profit of over Rs. 5,000 but not exceeding Rs. 10,000; 35 earn from Rs. 10,001 to Rs. 15,000; 25 from Rs. 15,001 to Rs. 20,000; 40 from Rs. 20,001 to Rs. 25,000; 50 from Rs. 25,001 to Rs. 30,000; 60 from Rs. 30,001 to Rs. 35,000; 40 from Rs. 35,001, to Rs. 40,000; 50 from Rs. 40,001 to Rs. 45,000; and 69 from Rs. 45,001 to Rs. 50,000, the subjoined is the

Frequency Table that would be made out from these crude data :—

Groups	Classes	Class Marks	Frequencies	Cumulative Frequencies
1.	—5,000	...	20	20
2.	5,001—10,000	7,500	30	50
3.	10,001—15,000	12,500	35	85
4.	15,001—20,000	17,500	25	110
5.	20,001—25,000	22,500	40	150
6.	25,001—30,000	27,500	50	200
7.	30,001—35,000	32,500	60	260
8.	35,001—40,000	37,500	40	300
9.	40,001—45,000	42,500	50	350
10.	45,001—50,000	47,500	69	419

Here the significance of certain terms used in connection with a Frequency Distribution should be explained. The difference between the highest profits in the class below the one in question and the lowest class above is known as *Class Interval* (also known as *Interpolation*). In other words, the class interval is the interval which sets bounds to each class of the frequency distribution. By *Class Mark or Number* is meant a value (generally the arithmetic mean of the class limits) which serves to designate the class.

The whole of the class range (that is to say, from Rs. 5,000 profit to Rs. 50,000 profit) is known as an *Array*, and the middle or the

central item in the array which most closely correspond to the magnitudes of all other examples in the array is known as the *Median*. The median is located half-way down the array. An array differs from a series in that whereas an array is an arrangement on the basis of magnitudes, the series is not necessarily such. It is rather an arrangement in logical sequences.

When an array is divided into four equal parts each such division is known as a *Quartile*. The quartile between the lower extreme and the median is called the *Lower Quartile*, and that between the median and the upper extreme is known as the *Upper Quartile*. Briefly stated the second and the third quartiles are respectively called the Lower and Upper Quartiles.

As will be perceived from the preceding table, a frequency table contains distinct columns for the recording of the serial number of the groups, classes, class marks and frequencies of each group. In the last column headed Cumulative Frequencies, the frequencies of the preceding classes have been cumulated, that is to say, superadded at each stage, and when this column appears in a frequency table, we call it a 'Cumulative Frequency Distribution Table.'

In the study of statistical problems a distribution may be found to be either of the *Symmetrical* or *Asymmetrical* type. It is said to be of the symmetrical type when the same number of observations or frequencies are found to be distributed at the same linear distance on either side of the midpoint of the frequencies, as illustrated in the following table :—

Symmetrical Distribution of Profits Made by 114 Companies*

Profit Class	Frequency
—5,000	2
5,001—10,000	15
10,001—15,000	25
15,001—20,000	30
20,001—25,000	25
25,001—30,000	15
30,001—35,000	2
Total...	114

Further a symmetrical distribution is called to be of the *Normal* type when the observations or frequencies are found to vary at a rate that would be *theoretically* expected in an infinite number of trials. Such a theoretical conception of its occurrence is derived from the mathematical law relating to the probable occurrence of chance phenomena in accordance with

*Hypothetical data.

binomial expansion. It simply means that "the values of a variate in a large number of cases tend to be distributed uniformly according to the mathematical law about the value that occurs the greatest number of times." The following is an example of *Normal Frequency Distribution* (constituting the binomial expansion of the tenth power) :—

Normal Frequency Distribution of Average Wages Earned by Workers in a Certain Factory*

Wages in Rs.	No. of Workers getting it
50	1
55	10
60	45
65	120
70	210
75	252
80	210
85	120
90	45
95	10
100	1
Total...	1,024

Under actual conditions, however, normal frequencies are seldom met with when dealing with economic problems. Rather, as a point of fact, in the majority of instances, a representative sample of economic data will show positive

*Hypothetical data.

distortion or skewness (see Ch. VI), and when normal frequency is noticed in such a distribution, unrepresentativeness of the sample is to be justifiably suspected and should not, therefore, be relied upon as a basis for judgment relating to the characteristics of the group.

A distribution would be called *Asymmetrical* when there are unequal number of items or frequencies lying to the left and right of the mode (see Ch. V).

CHAPTER IV

VISUALIZING THE DATA

One of the most familiar ways in which statistics can be presented to make them suitable for proper appreciation is by means of *Graphs.* To the businessman they are particularly helpful in presenting an effective visual picture of the trend of sales and purchases, price fluctuations, gross profits and expenses, turnover and net profits, as also various kinds of records (for instance, those relating to the various departments, factories, finance and costing, etc.) Their chief importance to the statistician, however, rests on the fact that they facilitate the preliminary examination of most data by bringing readily to the eye the salient features of the tabulated results of a compilation.

Graphs are constructed by plotting the statistical data on sectional papers (available in various scales) containing mutually perpendicular intersecting straight lines called *Axes.* The horizontal axis is usually called the *x-axis,* the vertical axis the *y-axis,* and their point of intersection as the *Origin.* In plotting a graph it is necessary to know only two points, and to join them by a straight line. The *point* is expressed by the symbol (a, b), where a

indicates the distance along the *x-axis* and *b* the distance along the *y-axis*.

The statistician should never omit to label every graph with a short and appropriate description and to mark the scales along the axis, for it must be remembered that a graph, however neat, is quite worthless for any practical purpose, unless it is provided with a label and scales.

Representation of Time Series

When time element exists in the data, we generally use the horizontal scale for the time element and the vertical scale for the magnitude of the items. A series of points at equal intervals are marked on the horizontal base line corresponding to a series of years, or months or whatever time interval is involved. The variable quantities are represented by vertical ordinates erected at points to the horizontal scale corresponding to the particular instants of time to which the statistics relate. To represent them in the form of a graph, all that is necessary is to mark only the end of the ordinate by means of a point on the paper, and then connect the consecutive points by means of straight lines. The result will be a curve. It is clear that any rise or fall in the magnitude

of the items we seek to illustrate will be reflected by a corresponding rise or fall of the curve in the graph. As the continuity of the curve can be extended with the entry of new time elements into the matter, such a curve is of great importance and usefulness in visualizing at a glance the progress of any business activity. Then again, different curves (either for different series of the same units or for two or more series of different units) can be plotted on the same chart, and so the businessman can at one glance compare the progress of one thing with the other, or different aspects of the same thing. Further, if the periodic moving average is as well plotted on the same chart it should enable the businessman to perceive whether the short term fluctuations have any permanent effect upon the solidarity or otherwise of the business.

When the vertical scale proves large enough to be consistent with the size of the paper, we eliminate a part of the diagram and note it by observing that the zero on the vertical scale does not coincide with the horizontal base line on which the time element is shown. But as diagrams are mainly constructed to help others in appreciating some statistical facts,

so it would be futile to represent several time series on the same chart, if the graphs nearly superimpose on one another or they cross and recross. When two or more series of figures belonging to different units are plotted on the same chart, the need arises of more than one vertical scale, but this difficulty can be easily obviated by sacrificing the units involved and resorting to percentages instead. When percentages are thus plotted only the relative figures are shown in the graph, instead of actual sizes of the figures, and, therefore, the diagram cannot pretend to show properly the original table. It should be noted that the basic figure in such series is 100 per cent, and this is emphasized by drawing a horizontal line through 100 per cent on the vertical scale.

When rises and falls are in their actual magnitudes plotted on the chart it is called the *Natural Scale Method.* The one great demerit of such a method is that it does not enable the businessman to know what ratio the fluctuations of one period bear to the other, for rises and falls of equal magnitudes are shown by the same vertical distance. For instance, if the fluctuations in one period be from Rs. 4 to Rs. 8 and in another period from

Rs. 50 to Rs. 54, the curve would move up the verticle scale exactly the same distance in space, yet the percentage of increase in the first period was 100 per cent and in the second period only 8 per cent. When the natural scale is used the base line should represent the zero, otherwise the true prespective of the rise and fall would be lost.

When however we wish to avoid the calculation of percentages, and like to show the ratios of the rise and fall instead of the rises and falls themselves absolutely, we plot them in what is known as the *Logarithmic Scale.* "To construct a Logarithmic Scale, we first find the logarithms of the numbers we desire to plot and divide our scale into such a number of equal divisions as will allow all the logarithmic numbers progressing in a uniform manner. The logarithms of the actual numbers are then plotted, instead of the numbers themselves." It should, however, be clearly borne in mind that such a chart merely shows the ratios of rises and falls, and not anything else. The chief thing to remember in this connection is that whereas in a natural scale graph "the same length on the paper in any part of the scale is equivalent to the same

number of units," " in a ratio scale this is not the case the length of the interval between two values on the scale is proportional to that between these two values." As a matter of fact, " consecutive points on the scale corresponding to consecutive integars get closer and closer together as we increase the numbers from 1."

A comparison of the Logarithmic Chart with the Natural Scale Chart often reveals much useful information. When the Logarithmic Scale is used, there should be no base line, otherwise fallacious conclusions will follow. Again, " there is no zero on logarithmic paper, as 'log O' is an indefinitely large negative number " (Printed Logarithmic Papers can be obtained from the market). The Logarithmic Scale is appropriately used for representing a " series involving different units or a series of figures which change fundamentally as time goes on, increasing or decreasing at a great rate."

Representation of Frequency Distribution

In plotting a Frequency Distribution we take the *x-axis* for the representation of the class marks and *y-axis* for the representation

of the frequencies. But since each frequency is an integar and includes all the individuals within the class interval to which it applies, the frequency is customarily represented, not by a single ordinate, but by a rectangle whose base is the class interval and whose length is equal numerically to the value of the frequency. The diagram formed by the frequency rectangles is called a *Histogram*, as distinguished from *Historigram* which is the representation of a historical or time series. From a glance at such a representation it would at once be clear that if the number of class intervals be increased, that is to say, if the divergence in the magnitudes of the class interval be smaller, narrower and shorter the rectangular blocks would be, so that if the narrowing process be continued sufficiently far, a smooth curve would result in place of rectangular blocks. The natural corollary of this is that the rectangular blocks do not afford us a correct view of the situation. This difficulty can however be obviated by the construction of a *Frequency Polygon* by joining together the outer ends of the base of the diagram right through the middle points at the top of each of the rectangles. Such a polygon would approximately amount in area to that of the rectangles.

Normal Frequency Curve

When a polygon representing a normal distribution is smoothened (in such a process all irregularities are to be smoothened out) into a curve, the curve so formed is known as the *Normal Frequency Curve* (also known as the Gaussian Curve after Karl Friedrich Gauss). This curve should begin and end on the same base line. In the representation of all chance or natural phenomena it would be easy to construct such a curve by simple elimination of all accidental variations. But in the representation of business phenomena, this method is strewn with difficulty on account of the uneven distribution (that is to say, when the number of items falling into class above and below but located at an equal distance from the modal group are not approximately the same) of the data. In such a case, it is obvious that the sides of the bell-shaped curve will not be symmetrical, and there would be present a skewness. When, however, the data are evenly distributed and the curve is of a symmetrical form, the median will be found to be located within the modal group, for it would always bisect the area of the diagram. But when skewness is present, the

median will be shifted from its proper position and will fall outside the modal group.

The Ogive

When instead of plotting the individual group frequency, we plot the cumulative frequency (that is to say, by addition of successive frequencies) of a series of observations, the resultant curve is known as the *Ogive* or the *Cumulative Frequency Curve.* The ordinates of such a curve are formed from a given frequency distribution by the addition of successive frequencies. In this case also, the original ogive line is smoothened into a curve and this is very useful in locating the 'median,' 'quartiles,' 'deciles,' etc. If the vertical scale of the graph represents the cumulative frequency and the horizontal the magnitude of the items, then the magnitude of the median can be located by marking the middle number on the vertical scale, then drawing a horizontal line from there to the curve and dropping a vertical line from there to the horizontal base (axis)—the intersecting point (origin) at the base, will give the magnitude of the median item. The same method can also be applied to obtain the magnitude of any other particular item under review—but it should

be noted that the mode is not very easily located on the ogive.

Other Charts

Sometimes, for the representation of statistical data we also use the *Bar Charts* (vertical, horizontal, etc.), *Maps* (cross-hatch, multiple dots, etc.), *Block Diagrams*, *Square Diagrams* and *Circular Diagrams*, *Silhouette Charts*, etc.; but useful though these devices sometimes are, they can never approach in importance in the mathematical analysis of the data to the ordinary method of graphing described above.

CHAPTER V

DETERMINING THE CENTRAL TENDENCY

When dealing with a large mass of data, it becomes difficult for us to comprehend its characteristics properly. On account of the difficulty which is thus encountered in grasping a large mass of figures, it is convenient to use a statistical type or average which sums up briefly the central tendency of the mass. It is, in other words, a useful way of representing the various values of a variable by a single value.

Averages generally used in statistical analysis are : (1) *Arithmetic Mean,* (2) *Quadratic Mean or Root-Mean Square,* (3) *Geometric Mean,* (4) *Harmonic Mean,* (5) *Median, and* (6) ***Mode.***

Arithmetic Mean

The Arithmetic Mean is the most common type of average that is used in statistical analysis. It is obtained by dividing the sum-total of the values of a number of items by the

number of items itself. Thus, if 993 joint stock companies were formed in 1936, 1,175 in 1937, 986 in 1938, 996 in 1939 and 1,005 in 1940, then the annual average thereof will be calculated by adding together the figures of the five years (5,155) and dividing the same by 5. This would show 1,031 to be the annual average (or Arithmetic Mean).

Sometimes we adopt a different method for the calculation of the average, and particularly when the items are too many in number and are of *near value*. In applying this method, we assume an arbitrary figure and take that to be the average of the group. We then add together (algebrical summation) the deviations of each of the items from the assumed average, and divide the total by the number of items. The quotient thus obtained (called the *Correction Factor*) when added to the assumed average will give the true average. Thus in the above example let us assume 1,000 to be the average. Its deviations (as shown in columns 3 and 4 of the subjoined table) from the actual figures, summated algebrically give 155 which divided by 5 yields +31, and this added to the assumed average (1,000) gives us 1,031 which as we have already seen, is the true mean of the

series. This is illustrated in the following table :

Company Flotations in India (1936-40)

Year	No. of Companies	Deviations from the Assumed Mean	
		Minus	Plus
1936	993	7	
1937	1,175		175
1938	986	14	
1939	996	4	
1940	1,005		5
Total ...	5,155	25	180
Assumed Mean	1,000		
True Mean ...	1,031		

It necessarily follows from this that had the assumed average been the true average, then the sum of the deviations would have been zero. In other words, the arithmetic mean is the point from which the algebraic sum of the deviations is zero. Interpreted in simple language this means that if the plus deviations are summated in one column and the minus deviations in another, the summation will be of the same size, with a difference of zero, showing thereby that the mean is the point around which the deviations reach a minimum.

It should be noted that an Arithmetic Mean though advantageous in many respects

is nevertheless fallacious and inaccurate when it is unsupported by the actual figures used in the computation of such an average. Again, though it is useful in showing the average size of items in a series or array, it gives no hint as to the extent of variation in magnitudes between the highest and lowest figures. This will be clear from the following table which shows that though the range of variations in the magnitudes of items of the two series (in one case from 1·1 to 21·2 and in another from 0·9 to 13·8) is considerable, yet their mean is the same.

Average Yield Per Cent Per Annum from 10 Groups of Equities in 1938 & 1939.

Securities		1938	1939
Jutes	...	1·1	0·9
Coals	...	5·3	6·2
Railways	...	4·7	3·3
Cottons	...	3·3	5·4
Electrics	...	7·7	7·7
Minings	...	10·1	12·2
Engineerings	...	21·2	13·8
Teas	...	3·3	4·0
Banks & Insurance	...	6·7	6·8
Miscellaneous	...	2·1	5·2
Mean	...	6·55	6·55

Sometimes when it is desired to give full weight or importance to variations, the *weighted average* is used. The weighted average is obtained by multiplying each item of a series of quantities by the number of subjects con-

nected with it, these multiples being called "*weights*"; the sums of the products when divided by the sum of the weights, will give us the weighted average. Thus, if 100 things are bought at Rs. 4 each and 50 things at Rs. 5 each, then the weighted average thereof will be calculated as follows :

$$\frac{(100\times4) + (50\times5)}{(100 + 50)} = \text{Rs. } 4{\cdot}4 \text{ per thing}$$

In other words, whereas in the calculation of the simple mean we consider each of the items of the series to be of equal importance, in the weighted average we do not. An example will clarify the point. Suppose, for instance we are to find out the average yield per acre of jute for 1938. We tabulate the data as follows in the table below.

Acreage & Production of Jute in 1938
(000's Omitted)

Area		Acreage	Production	Yield per acre
West Bengal	...	287·2	4,280	14·90
North Bengal	...	603·1	9,640	15·98
East Bengal	...	1,276·3	20,955	16·41
Cooch Behar, Tippera States & Nepal	...	42·3	610	14·42
Assam	...	219·1	3,275	14·94
Bihar	...	445·0	4,480	10·06
Orissa	...	15·3	165	10·78
Total	...	2,888·3	43,405	97·49
Simple Mean		13·927		
Weighted Mean	...	15·027		

Here the simple mean is obtained by dividing the sum total of figures in column 4 by 7 (the number of areas under review), but in doing this we take no account of the fact that the individual average yields are associated with varying acreages and production in the several areas. These differences are properly set off by the weighted mean. In this case the weighted mean is calculated by dividing the total of column 3 (production) by the total of column 2 (acreage). It should, however, be noted in this connection that when different values are associated with like quantities, the weighted and the unweighted averages are the same. But if different values are associated with unlike quantities the weighted and the unweighted averages will not be the same. In the latter case, with every change or alteration in the quantity, there would be a corresponding change in the magnitude of the weighted average. It can further be stated that " when all the factors of a group or universe are present in a sample in the same proportions as they exist in the group from which the sample was taken, the so-called weighted average and the simple average are of the same size." Hence, there is the necessity of a sample being representative in character.

The weighted average is useful for obtaining the average cost of articles bought at different periods or in varying amounts, or at different prices. It is particularly useful where quantities are in evidence.

In dealing with business or economic data, we often use the *Moving Average*. This is obtained by omitting from the component series the earliest item and taking in its place the most recent one. This is very useful in showing the nature of fluctuations over a given period. This is illustrated in the following table :

Retail Price of Rice at Rungpur 1924-33

Year	Price per Md. Rs.	Moving Average	Progressive Average
1924 ...	8·4	.....	.....
1925 ...	8·3	8·35	8·35
1926 ...	8·3	8·30	8·33
1927 ...	8·5	8·40	8·37
1928 ...	9·1	8·80	8·52
1929 ...	7·2	8·15	8·30
1930 ...	7·2	7·20	8·14
1931 ...	4·4	5·80	7·67
1932 ...	3·2	3·80	7·18
1933 ...	3·1	3·15	6·77
Mean ...	6·77		

In measuring the growth of progress of a new business, however, we sometime use the *Progressive Average.* Unlike the moving average, in this case we do not omit the earliest figures, but calculate the average by simple inclusion of the newest or the most recent figure. So the difference between a moving average and a progressive average is that whereas in a moving average the earliest data are omitted, in a progressive average, on the other hand, all the data are taken into consideration. Its chief usefulness lies in measuring the fluctuations of an item over a period for which no representative moving average can be obtained. But with the lapse of time when it becomes possible to obtain a representative moving average, a progressive average should be discarded in favour of it. For, in such circumstances a progressive average is likely to be distorted or biassed by the earlier data.

The arithmetic mean of a *Frequency Distribution* (particularly of a continuous type) is calculated by multiplying the midpoints of the classes by the number of frequencies in each class, and then dividing the summation of the individual products by the total number of frequencies in the distribution. The reason for using the midpoints of the classes is that,

assuming them to be the mean of the respective classes, the algebraic sum of the deviations is zero.

When, however, the distribution is not of normal character, it is better to calculate the mean by multiplying each individual item by its respective value, summating the products thus obtained, and then dividing by the total number of items. This will give the weighted mean of a frequency distribution. When, however, the distribution is of the discrete type, as interest rates and farm mortgages, the mean is calculated by multiplying each interest rate by the value of the mortgage, and then by the number of frequencies, summating and dividing by the total values of all the mortgages.

The application of the method is exemplified in the following table :—

Class	Frequency	Col. 1 × 2
1	5	5
2	8	16
3	3	9
4	2	8
5	1	5
6	2	12
7	1	7
8	1	8
9	7	63
Total...	30	133

$$\text{Mean} = \frac{133}{30}$$

$$= 4{\cdot}4333.$$

In other words, the mean of a frequency series is the value obtained by dividing the summation of the products of frequency times the class or variable by the total number of frequencies.

Quadratic Mean

A mean that is frequently used for various kinds of statistical analysis is the *Quadratic Mean* or the *Root-Mean Square*. It is obtained by extraction of the square root of the arithmetic mean of the squares of the items contained in a series. Thus, if we are to find out the quadratic mean of 4 and 5, it would be calculated as follows :

$$\sqrt{\frac{4^2 + 5^2}{2}} = \sqrt{\frac{16 + 25}{2}} = \sqrt{20{\cdot}5} = 4{\cdot}528\,(Q.M.)$$

Because of its use in connection with the measurement of standard deviations, the root-mean-square ranks as one of the most important of statistical averages.

Geometric Mean

Sometimes one or two items in a component series may have such disparate magnitudes, that the arithmetic mean thereof cannot be said to represent truly the magnitudes involved. In such circumstances, the Geometric Mean is of great advantage. The *Geometric Mean*

of a set of n positive numbers is obtained by simple extraction of the nth root of their product. In the calculation of the geometric mean, the use of logarithms affords great facility. The logarithms of the geometric mean is the arithmetic mean of the logarithms of the items contained in a series. Thus the geometric mean of 4 and 5 will be worked out as follows :

$$G.\ M. = \sqrt[2]{4 \times 5} = 4{\cdot}472.$$

$$\text{Or } \log G.\ M. = \frac{\log 4 + \log 5}{2}$$

$$= \tfrac{1}{2} (\log 4 + \log 5) = 4{\cdot}472.$$

Harmonic Mean

The *Harmonic Mean* is obtained by dividing the total number by the sum of the reciprocals* of the items. Thus the harmonic mean of the example we have previously worked out will be calculated as follows :

$$\frac{2}{\frac{1}{4} + \frac{1}{5}} = \frac{2}{\frac{9}{20}} = 4{\cdot}44.$$

Theorem : " In a series of positive terms, the quadratic mean is greater than the arithmetic mean, the arithmetic mean is greater

*Reciprocals are expressions so related to another that their product is 1. Thus 1/5 is the reciprocal of 5.

than the geometric mean, which in its turn is greater than the harmonic mean, unless the terms are equal in which case the values of the four are identical."

Median

We have already seen in connection with frequency distributions that when we arrange a series of items side by side in order of their ascending magnitudes, the range is known as an array, and the middle or the central item in the array which most closely corresponds to the magnitudes of all other examples in the array, is known as the *Median*. The Median is located half-way down the array, and for determining its position the following formula is used :

$$\frac{n+1}{2}$$

It should be noted that if the number of items in the array be odd, then there is no difficulty in locating the Median. For example, in the frequency distribution quoted on page 16, the median is the 210th item. In examples of even number, however, the median is fixed midway between two middle items. For example, if the number of items in the frequency distribution on page 16 had been 420, then the

median would have been the 210·5th item. (In the above formula n denotes the number of items in the array).

For measuring the magnitude of the Median, the following simple formula in preference to others given in most books on Statistics may be used :

$$M = L + \frac{cp}{f}$$

Where M=the Median ; L=the lower limit of the class in which the median is located ; c=the class interval ; f=the frequency of the class in which the median occurs ; and p=the number of items that must be counted in the median class to determine where the median occurs.

Thus, with the help of the above formula we can compute the magnitude of the median in the frequency distribution on page 16 as follows :

$$M = 30{,}001 + \left(\frac{5000}{60} \times 10 \right) = 33{,}125.$$

Mode

When in an array we meet with a predominating group of the same or approximately the same magnitude, the predominant group is called the *Mode* or the *Norm*. It is indeed

the item or value which occurs the *greatest number of times*, or the point of the greatest density, or of predominant or most fashionable value. In graphical representation the curve would flatten out at the region of the modal group. In the case of a group that is represented by a continuous series the value is the abscissa of the maximum ordinate. The value of the Mode can be determined with the help of the following formula :

$$Mo = L + \frac{CF}{F + f}$$

Where Mo=the Mode ; L=the lower limit of the modal group ; F=the number of frequencies in the next higher class, or in the class immediately above the one in which the mode is located ; C=the class interval; and f=the number of frequencies in the next lower class, or in the class immediately below the one in which the mode is located.

The mode is less definite in position than the median. But as compared with the mean, whereas the mean may correspond to no reality, the Mode is precisely the number for which the most numerous instances can be found. The special feature of the mode is that it remains

unaffected by any extreme. It is an indication of the type from which others are diverging. *

Quartiles

The *Upper* and the *Lower Quartiles* can very easily be calculated with the aid of the following formulas :

$$\text{Upper Quartile} = \frac{3(n+1)}{4}$$

$$\text{Lower Quartile} = \frac{n+1}{4}$$

* In a normal distribution both the median and the mode would coincide with the mean.

CHAPTER VI

MEASUREMENT OF SCATTER

Useful though they are in conveying to us an idea as to the character of the mass, the averages, however, do not furnish us any information as to the way the different values of the variable deviate from them. The way in which the different values of the variable deviate from the average is known as the *Scatter* or *Dispersion.*

Dispersion or scatter is measured by taking into account the extent to which the items representing the values of a variable deviate *on an average* from a standard type or item like the average, the median and the mode. In statistical analysis we note and record both the absolute and relative dispersion of a trait or character.

Absolute deviation is measured by calculation of the mean deviation of a series. To measure the relative dispersion of a trait or the ratio of the dispersion to the standard type, we use the coefficient (the fraction of variation occurring in a group) of dispersion

for each of the group under review. This is obtained by dividing the absolute measure of dispersion used by that magnitude which has been selected as representative of the data under review and from which the deviations have been measured.

Mean Deviation

A very satisfactory measure of dispersion in many cases is the *Mean Deviation*. To calculate it, we first summate the deviations from the standard type, and then divide the sum total by the number of items under review. In other words, it is merely the simple average of the deviations (without any regard for the signs). The following table is illustrative of this :

Retail Price of Rice at Rungpur 1919-24

Year	Price per Md. Rs.	Deviations from the Mean (d). Mean = 7·25
1919	7·5	·25
1920	7·9	·65
1921	7·5	·25
1922	5·3	1·95
1923	6·9	·35
1924	8·4	1·15
Total ...	43·5	4·60

$$\delta = \frac{\Sigma d}{n}$$

$$= \frac{4{\cdot}60}{6} = 0{\cdot}766.$$

The mean deviation from the mode and the median are also calculated in the same way. If d be the deviations from the average, dM the deviations from the median and dMo the deviations from the mode, and n the number of observations, the formulas for the determination of the mean deviation from the average, the median and the mode will be respectively as follows: $\frac{\Sigma d}{n}$, $\frac{\Sigma dM}{n}$, and $\frac{\Sigma dMo}{n}$. The result thus obtained will give us the absolute measure of dispersion, and the three kinds of deviations are symbolically represented as follows: δ, δM, and δMo.

In the case of a frequency distribution, however, the mean deviation is calculated by "determining the difference between the mean and the midpoint of each class, signs ignored, multiplying this difference in each instance by the number of frequencies in the particular class and then summating and dividing by the total number of frequencies in the distribution."

Coefficient of Dispersion

When comparing the dispersion of one group of observations with another we use the *Coefficient of Dispersion*. This Coefficient is obtained by simply dividing the average

deviation by the mean. To express the coefficient of dispersion as a percentage multiply it by 100.

Standard Deviation

Whereas the mean deviation is a measure of the extent to which the items in a series deviate *on an average* from a standard type, the Standard Deviation, on the other hand, is a measure of the extent to which the items deviate from the simple average, *giving weight to extreme deviations.* It is the quadratic mean of the deviations from the arithmetic mean, and is sometimes called the *Root-Mean-Square Deviation.* To calculate it, we square the deviation of each item from the mean, summate the squares, divide the summation by the number of observations, and then extract the square root of it. Stated simply, it is merely the value obtained by extracting the square root of the average of the squares of the deviations from the simple arithmetic mean, and can be symbolically expressed as follows :

$$\sigma = \frac{\Sigma d^2}{n}$$

Where σ = standard deviation ; d^2 = square of the deviation from the mean ; and n = the number of observations.

The following table illustrates the calculation of the standard deviation of a series :—

Retail Price of Rice at Rungpur 1919-24

Year	Price per Md. Rs.	Deviations from the Mean (Mean = 7·25)	Sq. of Deviations
1919	7·5	·25	·0625
1920	7·9	·65	·4225
1921	7·5	·25	·0625
1922	5·3	1·95	3·8025
1923	6·9	·35	·1225
1924	8·4	1·15	1·3225
Total ...	43·5	4·60	5·7950

$$\sigma = \sqrt{\frac{5{\cdot}7950}{6}} = \sqrt{0{\cdot}9658} = 0{\cdot}982.$$

The standard deviation of a frequency series is calculated with the help of the following formula :

$$\frac{\Sigma f\,(d^2)}{n}$$

Where f=frequencies ; d=deviation of the class from the mean of the series ; and n=the number of observations.

The application of the above formula is shown in the table below :

Series	f	d	d2	(Mean of the Series = 80).
60	1	—20	400	$\sigma = \sqrt{\frac{1000}{10}} = \sqrt{100} = 10{\cdot}00$
70	2	—10	200	
80	3	0	0	
90	4	10	400	
Total ...	10	*	1,000	

Standard Deviation of Two Groups

The standard deviation of two groups of data is calculated with the method exemplified in the following table :

x Group I	dx (Dev. from Mean)	y Group II	dy (Dev. from Mean)	(x-y)	(x-y)— Mean of (x-y)	Sq. of dev. of (x-y) from its mean
9	—7	10	—4	1	—1·2	1·44
12	—4	10	—4	2	—0·2	0·04
14	—2	13	—1	1	—1·2	1·44
15	—1	15	1	0	—2·2	4·84
16	0	13	—1	3	0·3	0·64
18	2	14	0	4	1·8	3·24
19	3	14	0	5	2·8	7·84
20	4	19	5	1	—1·2	1·44
22	6	17	3	5	2·8	7·84

Where x=group I ; dx=deviation from the mean of group x ; y=Group II ; dy=deviation from the mean of group y.

Mean of Group I=16.

Mean of Group II=14.

Mean of Col. 5=2·2.

Total of Col. 7=33·60.

$$\sigma = \sqrt{\frac{33{\cdot}60}{10}} = 1{\cdot}83.$$

Coefficient of Variation

When the standard deviation is expressed as a percentage of the average, it is known

as the *Coefficient of Variation.* It is obtained by multiplying σ by 100 and dividing the product by simple arithmetic mean. The formula for it is :

$$V = 100\ \sigma\ /\text{m}.$$

Quartile Deviation

This measure of dispersion is calculated by use of the formula :

$$\frac{Qu - Q1}{2}$$

Where Qu=the magnitude of the upper quartile ; Q1=the magnitude of the lower quartile.

The Quartile Coefficient of Dispersion is obtained by application of the formula :

$$\frac{\frac{Qu - Q1}{2}}{\frac{Qu + Q1}{2}} = \frac{Qu - Q1}{Qu + Q1}.$$

Lorenz Curve

Devised by Dr. Lorenz this graphic method is used for the measurement of divergences from the average. Dr. Lorenz used it specially for the measurement of the distribution of wealth, and it " takes the form of a cumulative percentage curve, combining the percentage

of items under review with the percentage of wealth or other factor distributed among such items." It is an ideal method for comparing the distribution of profits over various groups of businesses.

Skewness

Skewness is the distortion of symmetry of dispersion of items in a group by any reversal on opposite side. This is of the same significance as saying that "skewness" is merely "distorted normal distribution."* The formula used for the calculation of the *Coefficient of Skewness* is as follows :

$$s = \frac{A - Mo}{\sigma}$$

Where s=the coefficient of skewness ; A=the arithmetic mean ; Mo=the magnitude of the mode; and σ = the standard deviation.

In other words, it is calculated by subtracting the mode from the mean, and dividing by the standard deviation. "This quotient or value thus obtained is an expression of how many standard deviations, or what part of a standard deviation, the mean deviates from the mode. Whenever the mode is greater than the mean there is a minus quantity

* A normal distribution has, therefore, a skewness of zero.

when it is subtracted from the mean, indicating that the distribution is negatively skewed. When the mean is greater than the mode there is positive skewness. By merely determining the difference between the mean and mode without subsequent division by any measure we have an expression in actual values that serves as an indication of skewness. If, for example, the value of 6 is the mean and the value of 4 is the mode it can be seen immediately that the mode is 2 less than the mean."

Calculation of Probable Error

Pertinent in this connection are the various methods adopted for calculation of the probable error of the various statistical constants as well as the coefficients. For instance, the *probable error of the mean* is calculated with the help of the following formula :

$$\text{p. e.} = \frac{0{\cdot}6745\ (\sigma)}{\sqrt{n}}$$

In other words, to calculate it we multiply the standard deviation by 0·6745 and then divide the product by the square root of the number of items. "The probable error of the mean is quite commonly used as a measure of reliability. Whenever the probable

error is no greater than one-sixth of the standard deviation the sample of data may be considered fairly reliable, and when it is no greater than one-twelfth of the standard deviation it is a safe deduction that the sample is sufficiently reliable for practical purposes. In chance or random selections the probabilities involved in sampling are such that in only one case in a hundred is there a chance of the arithmetic average being inaccurate to the extent of more than four times the size of the associated probable error. That is, if we were to select a sample at random and then calculate the mean and the probable error of the mean there is but one chance in a hundred that the true mean of the entire universe would differ from the mean of the sample to the extent of more than four times the probable error of the mean of the sample."

Probable error of the standard deviation is calculated with the help of the following formula :

$$\text{p. e. } \sigma. = \frac{0{\cdot}6745\,(\sigma)}{\sqrt{2n}}$$

In other words, to calculate it we multiply the standard deviation by 0·6745 and then

divide the product by the square root of two times the number of items.

The formula that may be adopted for the determination of the *probable error of a distribution* is as follows :

$$\text{p. e. d.} = 0{\cdot}6745\ (\sigma)$$

In other words, to calculate it we multiply the standard deviation by 0·6745. This is as good as saying that "in any other distributions of like or similar data the chance are even that one-half of the number of the number of items will fall within a range of the mean of the first distribution plus or minus 0·6745 of the standard deviation."

CHAPTER VII

INDEX NUMBERS

By an Index Number is meant a value expressing the percentage of change taking place in the characteristic property of a series of items of a time series as compared with the level (written as 100) at any given base date. They are very widely used in the business field, and are useful in showing the relative magnitudes (expressed as a percentage on a base period) of changing or changed conditions from time to time. Thus, index numbers are applied to the measurement of the general movement of prices, cost of living, wages, production, consumption, employment, etc. As applied to the measurement of price changes, they are particularly useful in showing the reasons for the fluctuations of prices over a certain period. When the change in the price of two commodities under nearly equal conditions are compared, the common factor in such a comparison is, of course, a change in the purchasing power of money. When

the indices in regard to two or several commodities show similar fluctuations, it is assumed that the change in their prices is due to the general factor of a change in the purchasing power of money, but when there is any sharp divergence in a particular group, it is assumed that besides the general factor extraneous factors or causes are perhaps at work. In other words, index numbers are very useful in enabling us to measure the size of any hidden factor which though not capable of direct measurement can, however, be measured by taking into consideration of quantities which are influenced by such a factor.

The method that is generally adopted for the framing of an Index Number is to select the average or the actual value or magnitude of an arbitrary period as the *Base* (or equated to 100), and calculate the values or magnitudes of the subsequent periods as percentages on same. As applied to price indices, the problem concerned is of the same significance as that of the comparison of the purchasing power of a rupee in one year with its purchasing power in another. The series of proportional numbers obtained constitute what is known as the *Fixed Base Numbers.*

Chain Base Numbers

Sometimes, however, we frame what are known as the *Chain Base Numbers* (also known as the *Link Index Numbers*). A different method is, of course, adopted for this. In this case, we do not take the value or magnitude of any arbitrarily selected period as the base, but the average or the actual price of immediately preceding period as the base (100) and calculate the Index Numbers of the period under review as a percentage on same. An advantage of the Chain Base Method is that it enables a comparison possible with modern conditions, whereas in case of a Fixed Base Index the results are likely to be distorted or nullified if the Base be too old. When the year to year movements are of more interest than the change over a longer period of time, this method is of definite advantage.

The following is an illustration of both Fixed Base Numbers and Chain Index Numbers

of First Grade Jute Prices in Calcutta during February to December 1941 :—

Index Numbers of Jute Prices

1941	Price Rs. as.	Fixed Base No. (July 1914= 100)	Chain Index No.
February ...	32-8	47	100
March ...	37-0	54	110
April ...	35-0	51	95
May ...	44-8	64	127
June ...	48-8	70	108
July ...	51-0	74	106
August ...	66-0	96	129
September ...	69-0	100	104
October ...	61-8	89	89
November ...	63-0	91	102
December ...	53-0	77	84

Composite Index Numbers

When we construct an Index Number with the values or magnitudes of two or more things or commodities in it, it is called a *Composite Index Number*. Such an index number may be either of the weighted or unweighted character. To obtain an unweighted composite index number we add together the values or magnitudes for the base year, then divide the summated total for each year by the

total for the base year, and then multiply by 100. To obtain a weighted composite index number, we take the weighted average or actual values for the base year, and proceed to calculate as above. The most familiar type of weighted composite index number is the *Cost of Living Index Number.* This is illustrated by the following table :

Cost of Living Index on the Basis of Retail Prices Prevailing on the 23rd September 1942

For working class in and around Calcutta

Items	Weightage	Pre-war Index	Present Index	Weightage Present Index
(*a*) **Foodstuff**—				
Rice ...	24	100	202	4,848
Atta & Flour	11·3	100	186	2,102
Dal ...	6·7	100	189	1,266
Ghee ...	7·7	100	165	1,271
Oil ...	5	100	131	655
Salt & Spices ...	5·3	100	207	1,097
Sugar ...	5	100	125	625
Tea ...	1	100	100	100
Milk ...	9	100	100	900
Vegetables & Fish ...	25	100	100	2,500
				15,364
			Index	154

Items	Weightage	Pre-war Index	Present Index	Weightage Present Index
(*b*) **Fuel & Lighting**—				
Kerosene Oil ...	35	100	171	5,985
Coal & Firewood ...	61	100	231	14,091
Matches ...	4	100	180	720
				20,796
			Index.....	208
(*c*) Coarse Cloth	100	100	...	198
(*d*) and (*e*) Miscellaneous & House Rent—			Constant.	
Composite Index				
(*a*) Food	52·5	100	154	8,085
(*b*) Fuel & Lighting ...	7·5	100	208	1,560
(*c*) Coarse Cloth	7	100	198	1,386
(*d*) Miscellaneous ...	19	100	100	1,900
(*e*) House Rent	14	100	100	1,400
				14,331
			Cost of Living Index ...	143

In constructing a Cost of Living Index as above, we first determine the class of people (*e.g.*, industrial workers, artisans, clerks, etc.) for which the index numbers are to be compiled. We then collect a reasonably adequate number of sufficiently accurate samples of family budgets from the class under review. The period chosen for such budgets being

one of normal conditions, that is to say, free from abnormally high or low prices. The proportion of expenditure on different articles or objects by an average family is then determined, and the retail price quotations of these articles are collected from standard trade journals or municipal gazettes or official publications. If the quotations are obtained weekly and the cost of living index number is to be calculated for the month, the weekly prices are averaged into monthly figures. To start with, these monthly figures would form the base or 100, and similarly calculated figures for subsequent months would be represented as percentages of the prices of the base period. These percentages or index numbers for the particular commodities are then multiplied by their respective " weights " which represent the relative importance (the proportion which expenditure on each article bears to the total expenditure on the group) of the respective articles in the family budget. Then as shown above, index numbers of various groups of articles are arrived at by process of summation. These index numbers are then multiplied by the weights or the relative importance of the various groups of items, and by summating them we arrive at the composite index

number which forms the Cost of Living Index for the period under review.

To arrive at a correct Cost of Living Index Number, we must be careful in regard to demarcating the class of people under enquiry, correct selection of representative articles entering into the cost of living of the class under review, collection of reliable price quotations, accurate assignment of weights, and the changes in demand of various articles or their prices in the period under consideration.

Indices of Industrial Activity

These are useful in showing the change in the industrial production of a country over a period of time. These index numbers are prepared with the help of output figures of the different industries. The one prepared and published monthly by the *Capital* of Calcutta has 1935 for its base year, and the following series (with the respective weights within brackets) for its components :

I. Industrial Production—Cotton Manufactures (9), Jute Manufactures (6), Steel Ingots (5), Pig Iron (8), Cement (5), Paper (3). *II. Mineral Production*—Coal (7), *III. Rail & River-borne Trade* (24). *IV. Financial Statistics*—Cheque Clearances (20). *V. Trade*

Foreign & Coastal—Exports (4), Imports (3). *VI. Shipping, Foreign & Coastal*—Tonnage entered (3), Tonnage cleared (3).

Indices of Business Conditions

To show changes in the business conditions of a country a wider range of data is required, and Professor Pigou selected the following series for a study of the changes in the business conditions of England :

(*i*) Unemployment percentage. (*ii*) Consumption of pig-iron. (*iii*) Prices in England. (*iv*) Rates of discount on 3 months' bills. (*v*) Volume of manufactured goods. (*vi*) Agricultural production. (*vii*) Yield per acre of nine principal crops. (*viii*) Index of production from mines. (*ix*) London Cheque Clearings. (*x*) Increase of Bank Credit. (*xi*) Credits outstanding. (*xii*) Annual increase in the aggregate money wage. (*xiii*) Rate of real wages. (*xiv*) General aggregate consumption. (*xv*) Proportion of Reserve to Liabilities of the Bank of England.

CHAPTER VIII

CORRELATION & PREDICTIVE EQUATIONS

It is often noticed in the business field that there is a distinct relationship between certain allied sets of phenomena. This relationship is of the nature of cause and effect. Thus, there is found to be such a relationship between decline in production and rise in prices, employment and wholesale commodity prices, great industrial activity and high prices of equities, and so forth. The mathematical theory by means of which these relationships are found and reduced to formula and number is known as *Correlation*. The theory of correlation implies two or more sets of variables—those that cause or influence certain changes, and those that are the results or effects of such changes. The series producing the causal factors (*e.g.*, the output of a commodity) are generally known as *independent* variables (because the changes produced by it are independent of the other), and the series of factors which are the results or effects (*e.g.*, the price of a commodity) as the *dependent* variables (because they are dependent for their changes upon the factors of the other series). These two sets of factors are respectively termed as x values

and y values. Consequently, if there exists a causal relationship between output and price, the output is referred to as x and the price as y. Again, when the changes in the associated sets of phenomena are in the same direction the correlation is called *positive* ; when, however, it is of an inverse nature it is called negative.

Correlation can be studied from two or more aspects of a thing. Thus, for instance, when we study the relationship between the acreage and price of a crop, we call it a study in the correlation of two factors. Again, when we study the correlation between the acreage, yield, and the price of a crop we call it a correlation of multiple factors.

Coefficient of Correlation

For numerical measurement of the actual magnitude or degree of correlation that exists between two or more associated sets of phenomena we calculate the *Coefficient of Correlation*. The Correlation Coefficient is calculated by such methods that when perfect correlationship exists between the factors under review it has a value of 1 (wherefore the sign " 1 " is used for this), and when there is no such correlation the sign " O " is used. When

the correlation is of positive character, that is when it shows the deviations of both the factors in the same direction a plus (+) sign is used as a prefix, and when it is negative that is to say when the trends are in opposite direction a minus (—) sign is used.

Under actual conditions however, we seldom find a correlation that measures as great as 1, so that the value of the coefficient or correlation (represented by the sign r) is generally expected to be some fractional part of 1. The higher the value of the numeral in the coefficient, the greater is the degree of correlation. Whether this will be + or — quantity would, of course depend upon the direction of correlation. In this connection, the following rules as given by W. I. King in his *Elements of Statistical Method* for the interpretation of coefficient are to be noted :

1. If the coefficient is less than the probable error there is no evidence whatsoever of correlation.
2. If coefficient is more than six times the size of the probable error the existence of correlationship is a perfect certainty.

3. When the probable error is relatively small, if coefficient is less than 0·30 the correlation cannot be considered at all marked.

4. If the probable error is relatively small, a coefficient above 0·50 indicates decided correlation.

Pearsonian Coefficient

Karl Pearson's method for the measurement of biological correlation is very useful as well for the measurement of long-term fluctuations in the business world. This is also known as the 'Product-Moment Correlation.' For this method we use the following formula :

$$r = \frac{\Sigma (x - A)(y - a)}{n \sigma x \sigma y}$$

Where r=the coefficient of correlation ; x= the independent variables ; A=the mean of the independent variables ; y=the dependent variables ; a=the mean of the dependent variables ; n=the number of items under review ; σ x=the standard deviation of the independent variables ; and σ y=the standard deviation of the dependent variables.

The application of this formula is illustrated by the following table :

Correlation Between Production & Price of Jute

Year	Production (x)	(x—A)	(x—A)^2	Price (y)	(y—a)	(y—a)^2	(x—A) (y-a)
1931 ...	102	14	196	41·00	9·00	81·00	126·0
1932 ...	66	—22	484	25·25	—6·75	45·56	148·5
1933 ...	88	0	0	24·25	—7·75	60·06	0
1934 ...	88	0	0	29·75	—2·25	5·06	0
1935 ...	98	10	100	37·75	5·75	33·06	57·5
1936 ...	86	—2	4	34·00	2·00	4·00	—4·0
Total ...			784			228·74	328·0
Standard Deviation			$\sigma x = 11{\cdot}40$		$\sigma y = 6{\cdot}24$		

$$r = \frac{\Sigma (x—A)(y—a)}{n \sigma x \sigma y} = \frac{328{\cdot}0}{6 \times 11{\cdot}40 \times 6{\cdot}24} = +0{\cdot}769$$

Concerned as it is mainly with the measurement of long-time relationship, the Pearsonian Coefficient would prove unsatisfactory if our interest lies in short-term changes. For instance, if we were examining two long-term series that show positive trend for the period as a whole and only year to year negative trends, the Pearsonian method would not take into account these negative trends for the shorter period.

However, with a little modification, the Pearsonian method can be adopted for the measurement of correlation of short-term fluctuations. In this connection, we do not use the deviations of the dependent and the independent variables from the arithmetic mean. Instead, we use the deviations of dependent and independent variables from the *Trend*. For this purpose, we calculate the moving averages of the Index Numbers of the two factors, and take the deviations of these factors from the mean of the indices as the basis for the standard deviation in each case. Then the Pearsonian formula is applied to it.

Probable Error of Coefficient

We have already seen that for the interpretation of r, it becomes necessary to calculate the probable error of the *coefficient of correlation*. This becomes particularly necessary when a large number of representative cases are taken into account. The formula employed for the calculation of the probable error of the coefficient of correlation is as follows :

$$\text{p. e. } r = \frac{0{\cdot}6745\,(1-r^2)}{\sqrt{n}}$$

Where p.e.r.=probable error of the coefficient of correlation ; r^2=square of the coefficient

of correlation ; n=total number of items used in the calculation of the coefficient.

In other words, we subtract the product of the squares of the coefficient of correlation from unity (1), and multiply it by 0·6745, then divide the result by the square root of the total number of items used in finding out the coefficient. (For values of $1-r^2$ consult *Tables for Statisticians and Biometricians.* Table VIII).

It may be remarked that the probable error of correlation ratio and the correlation index are also calculated from the same formula as above, the proper measure being substituted for r.

Time Lag

In the study of correlation it is often found that changes in one series are not immediately followed by changes in another. In other words, there is noticed a time lag of several months or even of a year or more, before the independent factors are found to exert their effects upon the dependent factors. In calculating the coefficient of correlation between such series "it is a good plan to plot each one separately and then determine the approximate time lag by shifting back and forth."

This can be easily done by holding the charts before a light and moving them back and forth until the position of greatest relationship is determined).

Another method of "shifting" the data is illustrated by the following table interpolated from Kuznet's *Variations in Industry & Trade* (National Bureau of Economic Research, 1938) giving the indexes of seasonal variation of production and shipments of pneumatic castings in U.S.A. for the years 1923-31 :-

Month	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Index-Shipments (X)	89	81	98	110	113	120	129	123	104	87	72	74
Index-Production (Y) ...	96	101	114	113	116	112	97	104	91	91	82	83

Shifting the indexes for the production one place to the right, we get the new series, as follows :—

X ...	89	81	98	110	113	120	129	123	104	87	72	74
Y ...	83	96	101	114	113	116	112	97	104	91	91	82

and so on. Again, shifting to the left we have the series:—

X ...	89	81	92	110	113	120	129	125	104	87	72	74
Y ...	101	114	113	116	112	97	104	91	91	82	83	96

Correlation Table

When dealing with a large number of items, it is advantageous to calculate the coefficient

of correlation (signified by the letter r) with the help of a table known as the Correlation Table. This is illustrated by the following table :

Index of Wholesale Commodity Prices (1919–32).

Index of Employment (1919—1932)	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	Frequencies of rows
110-114									4	4
105-109								7	2	9
100-104								5	9	14
95- 99							6	25	11	42
90- 94							4	13		17
85- 89						2	2			4
80- 84					1					1
75- 79			2	3	4					9
70- 74			4	1						5
65- 69		2	2							4
60- 64		3								3
55- 59	6	2								8
Frequencies of columns.	6	7	8	4	5	2	12	50	26	120

The above is an example of the construction of a correlation table for determining whether or not a net relationship exists between wholesale commodity prices and employment, using the United States Bureau of Labour Statistics index of wholesale commodity prices and the index of employment from 1919-1932. The first step in the construction of such a correlation table is to divide the range of the two variates into convenient divisions. Here the commodity prices range from 104·5 to 62·6 and the employment index from 110·9 to 55·2. It will be convenient to divide the range for both series as follows: for the former nine divisions of five units each and for the latter twelve divisions of five units each. The respective frequencies will then be grouped as above.

A convenient method that is usually followed for the calculation of the coefficient of correlation from a correlation table is shown below :—

(Calculation of Coefficient of Correlation from a C. Table)

Index of Wholesale Commodity Prices (1919–1932)

y										g	g. y.	g. y².	T.	T. y.
6									4	4	24	144	16	96
5								7	2	9	45	225	29	145
4								5	9	14	56	224	51	204
3							6	25	11	42	126	378	131	393
2							4	13		17	34	68	47	94
1						2	2			4	4	4	6	6
0					1					1	0	0	0	0
—1			2	3	4					9	—9	9	—7	7
—2			4	1						5	—10	20	—9	18
—3		2	2							4	—12	36	—10	30
—4		3								3	—12	48	—9	36
—5	6	2								8	—40	200	—30	150
f	6	7	8	4	5	2	12	50	26	120	206	1356	215	1179
x	—4	—3	—2	—1	0	1	2	3	4					
f. x	—24	—21	—16	—4	0	2	24	150	104	215				
f. x²	96	63	32	4	0	2	48	450	416	1111				
S.	—30	—28	—16	—5	—4	2	28	156	103	206				
S. x	120	84	32	5	0	2	56	468	412	1179				

Index of Employment (1919—1932)

N. B. :—x and y = class marks ; f = total frequencies of columns ; and g = frequencies of rows.

$$x - \text{Av. } x = \frac{215}{120} = 1{\cdot}7917,$$

$$y - \text{Av. } y = \frac{206}{120} = 1{\cdot}7167.$$

$$\sigma x = \sqrt{\frac{1111}{120} - (1{\cdot}7917)^2} = \sqrt{6{\cdot}0481} = 2{\cdot}4593.$$

$$\sigma y = \sqrt{\frac{1356}{120} - (1{\cdot}7167)^2} = \sqrt{8{\cdot}3519} = 2{\cdot}8901.$$

$$r = \frac{\frac{1179}{120} - (1{\cdot}7917)\,(1{\cdot}7167)}{(2{\cdot}4593)\,(2{\cdot}8901)} = \frac{6{\cdot}7493}{7{\cdot}1076} = 0{\cdot}9496$$

Mean of Several Values of *r*

To obtain the average or mean of several values of *r*, we first convert *r* into *z* and then multiply each *z* by *N*—3, where *N* is the number of pairs of original *r*. The products are then summated, and the total divided by the summation of (*N*—3)'s. The quotient gives the mean value of *z*. The mean of the correlation coefficients under review will then be obtained by conversion of *z* into *r*. This is illustrated by the following table :

Calculation of Mean of Several Values of r.

If the same two variables are correlated in three groups, the numbers in the groups and the values of r being as follows :

Group	N	r
I	13	0·30
II	38	0·40
III	43	0·35

then the mean values can be calculated as follows :—

r	z	$N-3$	$(N-3)z$
0·30	0·310	10	3·100
0·40	0·424	35	14·840
0·35	0·365	40	14·600
		85	32·540

For $z=0·382$, $r=0·364$. Hence the average correlation in the three groups is 0·382.

Partial Correlation

We have so far studied simple correlation or the extent of co-variation between two variables. But a look at the economic or business situation will convince any man that most of the economic phenomena are influenced by a variety of factors rather than by one alone. Price of jute, for instance, may be due not merely to the acreage planted, but

also to the yield per acre, which in its turn is due to a multitude of factors such as labour, temperature, rainfall, fertilizer and irrigation. Similarly, in the Stock Market it has been perceived that long-term bond prices " respond to such situations as (*i*) changes in the cost of living, since the public regards bonds as relatively undesirable in protracted periods of rising prices, and *vice versa*, (2) the earnings applicable to the interest charges, the bond price responding to variations in earnings which threaten or fortify the coupon payments, and (3) other interest rates, which influence the height at which bond prices will capitalize their coupon payments."

In the field of chemistry or any other science of similar nature, it is easier for two or more causal factors of a phenomena to be isolated and analyzed. Such isolation and analysis are not, however, possible in the field of commerce and industry, as the multitude of factors producing their effects upon a phenomena can neither be controlled nor the relative importance of these factors properly assessed. Thus, for instance, if we were to study the influence of variations in money rates on business conditions, we cannot for the purpose

of our experimentation, keep the money rates inert and stationary for a period of years and then proceed to measure what effects that would produce on commerce and industry. This lack of control over economic factors is, however, made good by the statistical methods of *Partial* and *Multiple Correlations*.

In the study of correlation of this type, we measure the correlation of one set of independent variables with another set of dependent variables keeping the effects of any other set of independent variables constant, fixed or eliminated. This kind of correlation technique is known as the *Partial Correlation*. When only three sets of variables are under review all that we are concerned with in the study of such correlation is to determine the correlation between x and y denoted by the symbol $r$xy between x and z denoted by the symbol $r$xz, and between y and z denoted by the symbol $r$yz. If, for instance, the influence of z is eliminated or its effects or influence kept fixed and constant, then the correlation between x and y will be calculated with the help of the formula given below :

$$r\text{xy.z} = \frac{r\text{xy} - r\text{xz}.\ r\text{yz}}{\sqrt{(1 - r^2\,\text{xz})\ (1 - r^2\,\text{yz})}}$$

As explained above, read $r$xy.z as 'the correlation between x and y with the relation of z eliminated or considered fixed and constant.' This technique can as well be applied to four or more variables, and the formula to be used in that case will stand as follows :

$$r\text{xy.zm} = \frac{r\text{xy. z} - r\text{xm. z. } r\text{ym.z}}{\sqrt{(1 - r^2\,\text{xm. z})\ (1 - r^2\ \text{ym.z})}}$$

Here m stands for the fourth set of variables, and the problem involved is that of correlation between x and y keeping both z and m constant, and it requires the calculation of three other partial correlations, each keeping only one variable constant.

Calculation of Partial Correlation

If, for instance, we have three sets of data x, y and z under examination, and the correlation of xy, xz and yz are respectively as follows :-

$$r\text{xy} = 0{\cdot}50$$
$$r\text{xz} = 0{\cdot}40$$
$$r\text{yz} = 0{\cdot}60$$

then the correlation between x and y, keeping z constant will be calculated as follows :

$$r\text{xy.z} = \frac{0{\cdot}50 - (0{\cdot}40 \times 0{\cdot}60)}{\sqrt{(1 - 0{\cdot}16)\ (1 - 0{\cdot}36)}} =$$

$$\frac{0{\cdot}50 - 0{\cdot}24}{\sqrt{0{\cdot}84 \times 0{\cdot}64}} = \frac{0{\cdot}26}{0{\cdot}7328} = 0{\cdot}35.$$

Multiple Correlation

In the study of partial correlation our problem was to determine the relationship between two variables x and y, keeping the third variable z constant. In multiple correlation, on the other hand, our problem is to measure the combined effect of two or more independent variables upon the dependent variable. Calculation of the coefficient of multiple correlation becomes easier if the following formula is resorted to :

$$1-r^2\,x\,(yz)=(1-r^2\,xy)\,(1-r^2\,xz.\,y).$$

After finding out the value of $1-r^2\,x(yz)$, we simply subtract the result from 1 and get the value of $r^2\,x(yz)$, and then by extracting the square root of same, we get the value of $r x(yz)$ which expresses the coefficient of multiple correlation.

(It should be noted again in this connection that a table giving the values of $(1-r^2)$ for all values of r to 3 places of decimals may be found in *Tables for Statisticians and Biometricians*, and also in J. R. Miner's *Tables of* $\sqrt{1-r^2}$ *and* $1-r^2$ *for use in Partial Correlation & Trigonometry*, Baltimore, 1922, pp. 49).

Probable Error of Partial & Multiple Correlation

The significance of the partial and multiple correlation coefficients can be easily determined by means of their probable errors, which are symbolically expressed by the following formulas :

Probable Error of three variables :—

$$\text{p. e. of } r\text{xy. z} = \frac{0{\cdot}6745\ (1-r^2\ \text{xy.z})}{\sqrt{n}}$$

Probable error of four or more variables :—

$$\text{p. e. of } r\text{xy. zm} = \frac{0{\cdot}6745\ (1-r^2\ \text{xy. zm})}{\sqrt{n}}$$

Probable error of multiple correlation :—

$$\text{p. e. of } r\text{x (yzm)} = \frac{0{\cdot}6745\ \ (1-r^2\text{x(yzm)}}{\sqrt{n}}$$

In all the three cases n is the total number of cases used in the problem concerned.

Coefficients of Regression

Associated with correlation is the study of *Regression.* The term *Regression* is borrowed from biology. As applied to economic or business statistics, it means the tendency exhibited by either of correlated series to revert or regress toward its characteristic type. The study of regression is useful for predictive

purposes and the degree of relationship that exists between correlated items in the two series having a tendency toward reversion or regression, is shown by the line of regression which indicates the law of change in the mean of one variable for unit change in the other and if the line is straight the regression is said to be linear.

To calculate the ordinates of the regression line, the simple arithmetic mean of each of the two series are determined, as well as the coefficient of correlation between the corresponding items in the two series. The angles these lines make to the horizontal and vertical respectively are measured by the expressions :

$$r\frac{\sigma y}{\sigma x} \quad \text{and} \quad r\frac{\sigma x}{\sigma y},$$

and these are called the *coefficient of regression.* The former of these expressions means the regression coefficient of y on x, and the latter that of x on y. The values of y, the dependent most likely to be associated with given values of x, the independent, are obtained by employing the ordinary formula for straight line relationships,

$$Y = a + bx,$$

in which y is the ordinate of the straight line

of average relationship, a the arithmetic average, b the coefficient of regression and x the distance on the x-axis from the point where the mean of the x items coincides with the mean of the y items. The standard deviation is calculated either by squaring the deviations from the mean, summating, and dividing by the number of deviations, and extracting the square root, or by squaring the actual items, summating, dividing by the number of items, subtracting the square of the mean, and then extracting the square root.

In this connection, the significance of the equations

$$b = r\frac{\sigma y}{\sigma x}, \quad \text{and } b = r\frac{\sigma x}{\sigma y}$$

should be carefully noted. They are not interchangeable, as one might confuse them to be. In the first equation, it means that for each change of one unit in the "x" or independent variables, from the average change, there is a corresponding opposite change in the "y" or the dependent variables of a specified per cent (represented by the regression coefficient) of one unit. It gives the slope of the line of highest liner relationship between the two series. In the second equation, it

means that for each change of one unit in the "y" or dependent variables, there is a corresponding opposite change in the "x" or independent variables of a specified per cent (represented by the particular regression coefficient in this case) of one unit from the average year-to-year change.

In the example cited in connection with the calculation of Pearsonian Coefficient of Correlation on page 70, the coefficient of regression of "y" on "x," for instance, will be calculated as follows :—

$$b = r\frac{\sigma y}{\sigma x} = 0{\cdot}769\ \frac{6{\cdot}24}{11{\cdot}40} = 0{\cdot}421.$$

Predictive Equations

If it be realized that the line of regression is a means for determining the most probable value of the dependent variable when the size or magnitude of the independent variables are known, it would then be very easy to follow that the regression equation can as well, be appropriately used for the purpose of predicting the future trend of any dependent series of variables. But a question that may be pertinently asked in this connection is whether the actual conditions will coincide with the ordinate of a calculated trend of relationship

or the ordinate of the regression line ? There may be as a point of fact, considerable deviations from the ordinate of such calculated trend of relationship. With the help of the formula employed for the calculation of the coefficient of regression (which is an essential factor in predictive equation) we can indeed improve the value of the estimates, if the following formula be used :

$$Y = Ay - bAx + bx$$

For instance, in a problem in the estimate of the future yield of a crop on the basis of current price, we may substitute the following values for the symbols used in the above equation :

Y = the percentage change in acreage.

Ay = the average of percentage change in acreage.

b = the coefficient of regression of acreage on price.

x = the percentage change in price.

Ax = the average of percentage change in price.

(For b use the formula $r \frac{\sigma y}{\sigma x}$).

In other words, we simply subtract from the average percentage changes in the acreage the product of the coefficient of regression and the average of percentage change in the prices and the product of the coefficient regression and the percentage change in price. The method is illustrated in the subjoined table :

Year	Actual Acreage	% Change of Acreage	Mid-December Price Rs.	% Change of Price	Estimated % Change in Acreage $Y = Ay - bAx + bx$
1927 ...	36·30	24·04	63·50	3·24	1·65
1928 ...	33·71	—7·10	66·50	4·72	1·32
1929 ...	31·31	—7·11	54·50	—18·00	—1·60
1930 ...	33·17	5·92	29·75	—45·41	—7·63
1931 ...	34·86	5·10	41·00	37·81	5·96
1932 ...	18·62	—46·59	25·50	38·04	6·00
1933 ...	21·43	15·65	24·25	—4·89	—1·29
1934 ...	25·18	17·47	29·75	22·67	2·62
1935 ...	26·11	3·71	37·75	26·88	3·55
1936 ...	21·81	—16·43	34·75	—7·95	—0·61
1937 ...	28·22	29·42	34·25	—1·44	—2·05
1938 ...	31·66	12·15	33·75	—1·45	—2·04
Total ...	342·38	36·43	475·25	54·22	
Mean ...		3·35		4·52	

Our problem here is to estimate the acreage of jute crop for 1939, and be it noted here for that year alone and not for any other year in the series. In the last column are shown the figures of the estimated percentage changes in acreage obtained with the help of the predictive equation $Y = Ay - bAx + bx$. To apply this equation we had to calculate the coefficient of regression of jute acreage from 1927 to 1938 on that of mid-December prices from 1926 to 1938 (the price for 1926 was Rs. 61·50). The coefficient of regression in this case is 0·22. This is the value for "b" in the equation. The value for Ay is the mean of column 3 in the table above, and that for Ax is the mean of column 5. Multiplying this mean of percentage change by the coefficient of year to year percentage change in price as shown in column 5. Once we have thus found out the estimated percentage change in acreage (the figures relate to the next year) it is now easier to work out the actual probable acreage by simple multiplication of the acreage of any year immediately preceding the year for which the acreage is to be estimated by the percentage change and then adding or subtracting the product, as the case may

be. This is illustrated by the following table :—

Year	Estimated Change in Acreage	Estimated Total Acreage	Actual Acreage
1928 ...	1·65	36·89	33·71
1929 ...	1·32	34·15	31·31
1930 ...	—1·60	30·81	33·17
1931 ...	—7·63	30·64	34·86
1932 ...	5·96	36·93	18·62
1933 ...	6·00	18·73	21·43
1934 ...	—1·29	21·17	25·18
1935 ...	2·62	25·75	26·11
1936 ...	3·55	27·05	21·81
1937 ...	—0·61	21·68	28·22
1938 ...	—2·05	27·64	31·66
1939 ...	—2·04	31·02	31·60

Thus the estimated total acreage for 1939 according to the predictive equation $Y = Ay - bAx + bx$ is 31·02 lakhs acreage as against the actual acreage of 31·60 lakhs acres, showing thus a variation of only 1·8 per cent from the actual.

The slight variations that are noticeable between estimates and actual yields are indeed negligible for the purpose. It should be noted that where multiple correlation comes into the problem, the value of the predictive equations becomes doubtful. In this connection Harper has very rightly observed : " The value

of multiple regression in forecasting is probably sometimes over-emphasized, particularly in regard to strictly economic data."

Standard Error of Estimates

You must have noted by this time that although it may not invariably be the case, yet generally speaking the extent of accuracy of any estimates or predictions " vary directly in accordance with the magnitude of the coefficient of correlation." For instance, if there be " abnormally large variables in one series not compensated for in another we may obtain a high expression of relationship which is not a true index of actual cause and effect." Where therefore, there is not much consistency of relationship, the *error of estimates* is used to determine the " extent to which estimated or predicted values deviate from actual values." "To calculate it, we simply extract the square root of the mean-square deviation of estimated or predicted values from the actual." " The standard error thus calculated may be taken to indicate that if the curve of estimates were normal a distance equal to the standard error measured from either side of the mean of estimates would include approximately 68·26 per cent of all

the estimated values. When the standard error of estimate is larger than the standard deviation of actual values the estimating or predictive equation cannot be relied upon."

An alternative method for determining the error of estimates is by the *Coefficient of Alienation*. This is based upon " the mean of squared deviations of actual items from their arithmetic mean and on the mean of squared deviations of estimated or predicted values from the actual values, and it involves the calculation of both the standard error of estimate and the standard deviation." The formula employed for this may be stated as follows :—

$$ca=(\mathrm{SEy}^2)/(\mathrm{SDy}^2)$$

where ca=the coefficient of alienation ; SEy^2 =the square of the standard error of estimate ; and SDy^2=the square of the standard deviation of actual values.

" Since the standard deviation is merely the root-mean-square of the deviations from the arithmetic mean of the actual data, and since the standard error of estimate is simply the root-mean-square of the differences between the actual and estimated or predicted values, it necessarily follows that when the square of the standard error is divided by the square

of the standard deviation a quotient is obtained that represents the root-mean-square of the differences between the actual and estimated or predicted values in terms of the root-mean-square deviations of actual values from their arithmetic mean. Merely dividing the standard error by standard deviation would give a quotient of little or no significance. This is because both the standard error and the standard deviation are derived from the means of squared deviations, and not simply from the deviations as such. It is no more than a truism to state that squares of numbers do not increase in the same proportion as the magnitudes of the numbers themselves! The square of 12, for example, is 144, whereas the square of 17 is 289, and 289 is more than twice the size of 144, and although 17 less than one-half time the size of 12." (Harper, *Elements of Practical Statistics.*)

CHAPTER IX

BUSINESS FORECASTING

The basic theory of business forecasting rests on the belief that in normal circumstances, there is a definite cycle of ups and downs of business activity. Such ups and downs of business activity fall under four distinct categories :

I. Secular or long-period trends, caused by :

(*i*) growth of population,

(*ii*) improvements in methods of production,

(*iii*) exhaustion of natural resources,

(*iv*) decline in demand due to change in habits,

(*v*) competition of substitutes,

(*vi*) efficiency of management, and

(*vii*) change in the purchasing power of money.

II. Cyclical movements, caused by :

(*i*) maladjustment of economic activity,

(*ii*) psychology of business community,

(*iii*) monetary conditions, and

(*iv*) political conditions.

III. Seasonal variations, caused by :

(*i*) the difference in the length of month,

(*ii*) holidays, and

(*iii*) the difference in weather conditions.

IV. Accidental detriments, caused by :

(*i*) wars,

(*ii*) strikes and lock-outs,

(*iii*) natural catastrophes, *e. g.*, floods, earthquakes etc., and

(*iv*) changes in management or policy.

Of these four categories of ups and downs in the business field, no forecasting is possible in regard to accidental detriments—insurance being the only safeguard in the matter. Seasonal variations move with greater regularity than any other movements, and forecasts about them are easy with detailed weather and crop

reports. Scientific forecasting concerns itself only with the forecasting of cyclical movements, so that with a knowledge of such forecasts the effects of secular trend may be eliminated, and the intensity of trade cycle itself may be reduced and the range of price fluctuations lessened. Various control measures are adopted for this, the most familiar among them being the regulation of credit and currency policy. This is usually done through the central banking institution.

Relevent statistical series generally employed for the purposes of forecasting sequences or those of a trade cycle are the figures of production in relation to costs and stocks, bank clearings, bank deposits, railway goods traffic, electric power production, wholesale prices, employment and unemployment, export and import trade, pig iron production, business failures, wholesale and retail sales, stock exchange transactions, money rates and the cost of living index. In the selection of such data the following qualities of these should be kept in mind : representativeness, reliability, sensitiveness and frequency of publication. Besides, the information must be up-to-date.

Various series just named for the measurement of business activity are considered and examined not only as single series, but composite indices are also constructed for the purpose. In this connection Dr. J. H. Richardson offers a very pertinent note of warning : " Composite indexes are of value for broad general purpose, but it is essential to examine also the individual series upon which they are based. Much information of real importance may be concealed from view in a composite index. In fact, equal but opposite tendencies shown by two individual series will be cancelled when they are combined. Thus, a composite index may show the general state of trade to be the same now as a year ago. But this may be the result of a considerable increase of activity in some industries and of decline in others. Such a change may foreshadow the development of serious disequilibrium which will precipitate a crisis. This danger can be revealed only by a study of the individual series. The need for this study is generally recognized by forecasting services which give in full detail the individual series from which their composites are constructed. They are fully aware that a satisfactory review of the general state of trade can be made only by an examination

both of a composite index and of separate data indicating the situation of each of the important branches of business activity."

Here we must say something about the statistical methods to be applied in dealing with those statistical series. In the first place, for the convenience of comparison, index numbers are constructed of the relevant series. As trade cycle fluctuations are complicated by secular trends, these should be determined and eliminated. As we have shown below, the line of secular trend is calculated with the aid of the method of *least squares*, and with a chart drawn this can be easily eliminated. It can also be eliminated with the *moving averages*.

For the elimination of seasonal fluctuations the London and Cambridge Economic Service uses the following method. "Suppose the statistical series under consideration shows a figure for each month for twenty years. Then the average annual figure for the whole period is calculated and also the average of the twenty January figures, the average of the February figures, and so on. The average figure for each month is then expressed as a ratio of the annual average and these ratios are then

used for the elimination of seasonal variation from the actual figures.....The same process would be applied for this purpose to the figures for each month throughout the entire period covered by the statistics."

Other techniques generally adopted in connection with forecasting are the methods of standard deviation for the measurement of variability and that of correlation for the establishment of causal relationships between associated groups.

Determination of Secular Trend

Of vital interest undoubtedly to the businessman is the technique employed for the determination of the general inclination of a time series over a long period of time. This general inclination or tendency of a time series is known as the *Long-term* or *Secular Trend*. Various methods are employed for the fitting of a line or curve to this type of data.

When only an approximation of the general tendency of the series is desired, the best method appears to be the fitting of a trend by the free-hand method. This consists of drawing a line through a graph in such a way as to represent the general course of the series.

The fitting of the free-hand trend is sometimes made easier by the plotting of a three-year or five-year moving average as a stepping stone to the process. "The essential features to remember is that the first ordinate of the moving trend is always plotted for the mid-year of the series for which the average is calculated, or for any other mid-point that may be represented on the x-axis." It should be remarked that although the moving average has significance enough "in making approximations of the general movements in a series, and particularly in eliminating a large part of the cycle that is rather regular," yet it is "not satisfactory when there is no pronounced cycle and when the curve representing the actual values in the series shows sudden or marked changes in any direction that are not a part of the general cyclical movement or tendency."

There is also another method for determination of the secular trend known as the *Straight Line of Least Squares*. It is particularly useful in connection with a series that shows a general movement in one direction over a long period of time. This technique cannot, however, be applied to a series which having first shown an upward movement, later on moves downward

in a decided slope. The application of this technique is illustrated in the sub-joined table :

Straight Line Trend of Yield of Jute per Acre in India, 1921-1939

(Line of Least Squares)

Year	Yield per acre in lb. (Mean = 1413)	Deviation from the midpoint	Product of yield per acre & deviation from the midpoint (Co. 2 × Col. 3)	Sq. of Deviation from the midpoint (Sq. of Col. 3)	Ordinates of the straight line trend of least squares
	y	x	xy	x^2	Y
1921	1241	—9	—11,169	81	1557
1922	2080	—8	—16,640	64	1541
1923	1720	—7	—12,040	49	1515
1924	1601	—6	—9,606	36	1509
1925	1320	—5	—6,600	25	1493
1926	1400	—4	—5,600	16	1477
1927	1360	—3	—4,080	9	1461
1928	1321	—2	—2,642	4	1445
1929	1320	—1	—1,320	1	1429
1930	1281	0	0	0	1413
1931	1040	1	1,040	1	1397
1932	1400	2	2,800	4	1381
1933	1641	3	4,923	9	1365
1934	1360	4	5,440	16	1349
1935	1480	5	7,400	25	1333
1936	1420	6	8,520	36	1317
1937	1420	7	9,940	49	1301
1938	1241	8	2,928	64	1285
1939	1201	9	10,809	81	1279
Total ...		0	—8,897	570	—

The main point to be remembered in this connection is that the ordinates of the straight line trend of least squares are determined with the help of the formula :

$$Y=a+bx$$

Where a=the arithmetic mean of the series ; b=the slope of the line of least squares ; x=the deviation from the point of origin ; Y=the ordinate.

The slope of the line of least squares is calculated with the formula :

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

Where b=the slope of the line of least squares ; Σxy=summation of the product of each item in the series multiplied by the corresponding deviation from the midpoint ; Σx^2= summation of the squares of the deviations from the midpoint.

In the above example $b = \frac{\Sigma xy}{\Sigma x^2} = \frac{8{\cdot}897}{570} = -16.$

$Y=a+bx=1413+$ ($-16 \times$ deviation from the point of origin).

CHAPTER X

STATISTICAL REASONING

The interpretation of statistics or the drawing of inferences from numerical facts is a job for the expert. Statistical methods are indeed most dangerous devices in the hands of an inexpert. The necessity arises, therefore, for statistics being handled only by experts.

Blunders in statistical reasoning may be due to various causes. But the most common sources of error are the use of inaccurate or incomplete data, dishonest or misleading methods of presentation, false generalizations and faulty use of statistical methods. We have already seen that patient collection of facts or data constitutes the first most important step in all statistical investigations. For, no statistical results can be arrived at that are not already implicit in the data, and the accuracy of the former depends on that of the latter. In this connection we have noticed that one of the common statistical methods for the collection of data is Sampling. It is from the characteristics of the sample that we infer the characteristics of the *population*

or the bulk from which the sample is taken. A population consists of *individuals*, and although the individuals in a population vary, nevertheless they merge in the population and their individuality is lost. The formulas and laws describing the behaviour of populations as opposed to individuals are known as *Statistical Laws*. One of the fundamental laws warranting the reliability of sample data is the Law of Statistical Regularity based upon the mathematical Theory of Probability. But there may be errors of sampling, and 'the error in a sample result depends on the size of the sample, on the nature of the bulk being sampled (particularly on the variation within it) and the way in which the sample is taken.' Broadly speaking, other things being equal, such sampling errors are proportional to the amount of variation in the population. The biggest error, therefore, decreases as the size of the sample is increased. Errors of a sample may be either a Standard Error or an Error of Bias. Mathematically speaking, the standard error is inversely proportional to the square root of the number in the sample. On the other hand, error of bias is displayed in representative sampling of the type used in Gallup polls of public opinion, and it is necessary

to use very elaborate sampling methods to avoid errors of this kind. As an example of such " biassed " use of statistics, we may cite here the results of Gallup Polls on American Presidential election of 1944, held on August 5, that is to say, thirteen weeks before the actual elections. *Fortune* survey showed 52·5 per cent score in favour of Roosevelt, 43·9 per cent in favour of Dewey and 3·6 per cent Don't knows. This was found to conform more approximately to the actual results of the election held in next November. Simultaneously, Pollster George Gallup also had a score sheet, which revealed 51·3 per cent votes in favour of Dewey and 48·7 per cent votes in favour of Roosevelt —which was evidently based on a biassed sample and, therefore, yielded a wrong result.

Use of irrelevant or incomplete data and the drawing of false generalizations therefrom is a common trick of propagandists and commercial advertisers. Some interesting examples of insidious presentation of statistical data to slip across dubious arguments to the public are cited by Mr. L. H. C. Tippet in his recent work on *Statistics* published by the Oxford University Press. The following advertisement appeared in 1931 : " It is men

of exceptional experience who are buying X..... cars today. 87 per cent of X.....cars today are bought by men who have owned six other makes of cars before." Mr. Tippet comments : "I suppose it is unlikely that as many as 87 per cent of all makes of cars are bought by such veterans as those mentioned in the advertisement, and the purchasers of X.....cars are probably exceptional, but they may be exceptional in their fickleness—and do the makers of the X.....cars wish us to believe that they do not get many repeat orders ? These are possible interpretation of the data." Another possible explanation is that X.....cars were not very low-priced, and statistically speaking salaries increase with age. Another example is extracted from a report on an enquiry instituted some years before the last war as to the effects of the use of oatmeal among children and in public institutions. The Report states : "In Manchester 2,333 children in all were questioned. In one school of 200, 84 per cent were regular users, and the teacher stated, judging from regularity of attendance, that those getting oatmeal were the most satisfactory. In a girls' school of 182 pupils, 33 had porridge, and the head-mistress reported : the majority of oat users

are strong children, well-nourished, and class work good. The non-users are not strong and more liable to take colds and infectious diseases, and class work only moderately good." On this Mr. Tippet comments: "Now the essential information in the above extract is the better health of children who use oatmeal, but this is given only in vague qualitative terms. No statistician would rely on such general impressions as are quoted. What were their sickness records? Moreover, the poorness of the data is covered over, doubtless unintentionally, by some very exact but irrelevant figures. It does not matter two hoots how many children were questioned, or how many took porridge. Without these figures, the data would be seen to be what they are—weak. Most people take milk with porridge, which might be extra to milk taken otherwise, and that might be the cause of the improved health. All we know, if we know anything from the data, is that oatmeal plus milk plus the condiments are good for health as compared with the food that is eaten as an alternative."

Something may now be said about the fallacies of statistical reasoning due to the limitations of statistical methods. For instance,

although there is no single statistical quantity more valuable than the average, yet statisticians are at great pains to stress the inadequacy of this constant. Professor Bowley for instance observes : " Of itself an arithmetical average is more likely to conceal than to disclose important facts ; it is of the nature of an abbreviation, and is often an excuse for laziness." In fact, the average fails to measure the important facts that arise from variation. As ' the strength of a chain is the strength of its weakest link, not that of the average link,' so ' when data are in the form of a time series and averages are taken over a long period of time, they are apt to conceal important changes in the trend.' Mr. Tippet remarks : " A development of great importance in applied statistics has taken place during the past decade or so, and the results form a recognition of the fact that variation is a composite quantity, resulting from the combined effects of a multitude of factors." This is particularly evidenced in wrong interpretation of Index Numbers, in the making of which averaging plays an important part. For instance, from a rise in price index it may be argued that there is inflation in the country. But the value of such an argument is dubious unless

it is recognised that multitudes of causes have their effects on the index numbers and all the various factors concerned are taken into account. The index number here merely reveals a change in relationship and does not prove a case. It has indeed been very rightly observed that "the use of statistical data to prove a case, in the sense of demonstrating it, is unscientific; but their use to prove a case in the old-fashioned sense of testing it is scientific and profitable. A statistical inquiry should be approached with a mind that is open but not empty."

Many faulty arguments also arise from the wrong interpretation of the Coefficient of Correlation. As an instance of such nonsense correlations, we may cite the fact "that the proportion of marriages solemnized in the Church of England and the death rate for the country have for many years been decreasing—there is a correlation between the two."

So much indeed for the "don'ts" that the statistician should beware of in handling statistical data. Let us now consider the "do's" that should be observed when drawing inferences from statistical data. The statistical method, it should always be borne in mind,

is not different from the general scientific method. As a matter of fact, it is part of the same method, and is based on the same fundamental ideas and processes. In other words, statistical reasoning is not different from any other kind of reasoning. And as it is in any other field of knowledge, the use of good working hypotheses is the most essential aspect of statistical reasoning. "The ability to formulate fruitful hypotheses and design experiments to test them is the quality of a first-rate scientist. In addition to this personal quality, habits of thought and even prejudice have their influence on the kinds of hypothesis that will be entertained. For this reason impartiality is essential; and an investigator is most likely to be impartial if he is disinterested in the issue of the inquiry. The investigator should not be narrow minded, and should be prepared to consider any reasonable alternative to the main hypotheses he favours, but he cannot afford to waste his time on unreasonable ones."

As a matter of fact, a statistician should not be dogmatic about his conclusions. He should indeed have a 'critical apparatus sufficiently well-developed and discriminative to

prevent an undue proportion of false conclusions being reached as a result of statistical inquiries.' He should take an impartial view of things and should not suffer himself to be stimulated by any emotional appeal. For instance in connection with the enquiry relating to the benefit of oatmeal at the Manchester schools cited by Tippet, any of the following hypotheses may be adopted : the benefit may be derived from (*a*) oatmeal alone, (*b*) milk alone, and (*c*) neither oatmeal nor milk. Then the value of these different hypotheses may be tested by measuring separately the health of children who took (*a*) oatmeal and milk, (*b*) oatmeal without milk, (*c*) milk alone, (*d*) oatmeal alone, and (*e*) neither milk nor oatmeal.

When in testing hypotheses it is found that the data are capable of satisfying several reasonable hypotheses, the need arises then of fresh data being collected before any discrimination can be made between them. In this connection it should, however, be noted that "the favourite hypothesis with which the statistician usually first examines data is that the observed variations and effects are due to random errors or to chance rather than to the operation of newly discovered causes."

Applied Statistics

Applied Statistics deal with the practical application of statistical rules and formulas to concrete subject matters like prices, production, wages, trade, etc. In recent times much applied statistical work has been done in the business and economic fields. In U. S. A., for instance, statistics play a large role in the development of general policy, management, production, forecasting of business and trade, separation of cyclical, seasonal, and random movements of business trend, estimating of the elasticity of demand, ascertaining of consumer markets and public opinion, formulation of sales and advertising policies, and last but not the least in the development of life insurance. In connection with such business and economic investigations statistica data are used either with a view to formulating new theories, or testing of existing theories, or providing a measure of quantities that emerge from economic analysis. Needless to observe, there is need on the part of such investigators not only to have expert knowledge of statistical methods, but of the technical aspects of the problems under inquiry.

CHAPTER XI

CALCULATION BY LOGARITHMS

Students working in a statistical laboratory have the advantage of many a mechanical device for their calculations. For instance, he has the Calculating Machines, the Slide Rules, the Card-sorting Machines, the Correlation Calculators, and so forth.

But as these devices are not readily accessible to the ordinary student he must, therefore, alternatively have some simpler equipments like Tables of Logarithms, Square Roots, Cube Roots, Reciprocals, etc. There are various editions of these in the market, and for ordinary practical purposes the statistician is advised to have a copy of *Four Figure Mathematical Tables* by Frank Castle (Macmillan) obtainable at the nominal price of a few annas from any book-seller. For more advanced purposes, they are advised to have Chambers' *Mathematical Tables, Mathematical Tables No. I* (published by Statistical Laboratory, Calcutta), *Tables for Statisticians and Biometricians* by Karl Pearson, Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals of*

all Integral Numbers up to 10,000 (E. & F. N. Spon).

In practical work calculation by logarithms saves much time. A Logarithmic Table or the Log Table as it is commonly called, is a collection of auxiliary numbers so devised that

(*i*) multiplication of common numbers can be performed by the addition of their logs.

(*ii*) division by their subtraction.

(*iii*) involution or raising of powers by their multiplication.

(*iv*) evolution or extraction of roots by their division.

Thus if x and y be two numbers, the log method can be symbolically expressed as follows :

(*i*) $\log (x \times y) = \log x + \log y.$

(*ii*) $\log x \div y = \log x - \log y.$

(*iii*) $\log x^n = n. \log x.$

(*iv*) $\log \sqrt[n]{x} = \frac{1}{n} \log x.$

This is based upon the simple algebrical Law of Indices by which we know that

$x^2 \times x^2 = x^{2+2} = x^4$, or $x^4 \div x^2 = x^{4-2} = x^2$,

and so forth. The integral part of a logarithm is called its *characteristic* and the decimal part is called the *mantissa*. Only the mantissa is found in the Log Table, and the characteristic is obtained by inspection in accordance with the following rule: *The characteristic of a number greater than unity is positive and is one less than the number of digits to the left of the decimal point; the characteristic of a number less than unity is negative and is greater by one than the number of zeros which follow the decimal point.*

A bar is usually put over a negative characteristic. Thus, the characteristic of

2134·0	is 3	·2134	is $\bar{1}$
213·4	is 2	·02134	is $\bar{2}$
21·34	is 1	·002134	is $\bar{3}$
2·134	is 0	·0002134	is $\bar{4}$
		and so forth.	

Thus the characteristic of the logarithms of numbers from 1 to 9 inclusive is 0. In like manner the characteristic of logarithms of numbers from 10 to 99 is 1, whereas the characteristic of logarithms of numbers from 100 to 999 is 2. The characteristic of logarithms of numbers from 1000 to 9999 is 3, and so on for all other numbers.

The mantissa may be expressed with a large number of digits, but for all practical purposes in statistical calculations a four-figure fraction will suffice, the last digits being rounded off in the same way that 21·337 per cent is expressed as 21·34 per cent.

Logarithms of numbers are found by first determining the characteristic and then looking up the mantissa. Let us, for instance, find the log 21·34. According to the above rule the characteristic will be 1, and to find the mantissa we refer to the Log Table. There to obtain the mantissa corresponding to 21·34, we look up the index opposite to 21 in the column headed 3. There we get ·3284. Then for the fourth significant figure we refer to the difference column headed 4 at the right hand side of the Table and get 8 which we add to ·3284. The mantissa is thus ·3292 and the required logarithm is therefore 1·3292. Where the difference column is not furnished, the fourth significant figure is determined by taking the difference between the mantissa of log 21·3 and log 21·4 multiplying the difference by 0·5 and then adding the product to the mantissa of log 21·3.

It should be remembered that the answer to any problem calculated with the aid of

logarithms, and to find the number (known as antilogarithm) corresponding to a given logarithm, it is but necessary to locate the mantissa of the logarithm in the table and then read off the number corresponding thereto. Suppose the logarithm is 1·00. Reading down in the table of logarithms we find the fraction. ·0000 is located opposite 10 in the column headed 0. The characteristic 1 tells us that there must be two digits in that part of the number preceding the decimal. Accordingly we record 10 as the whole number and ·00 as its fraction. The number corresponding to the logarithm is therefore 10·00.

As we have seen, to perform multiplication by logarithms, we add the logarithms of the multiplier and the multiplicand, and their sum is the logarithm of the product. Thus,

$$\begin{aligned} 21{\cdot}34 \times 213{\cdot}4 &= \log 21{\cdot}34 + \log 213{\cdot}4 \\ &= 1{\cdot}3292 + 2{\cdot}3292 = 3{\cdot}6584 \\ &= 3{\cdot}6584. \quad \text{Its antilog} \\ &= 4554 \text{ the required product.} \end{aligned}$$

To perform division by logarithms we subtract the logarithm of the diviser from the logarithm

of the dividend, and the remainder is the logarithm of the quotient. Thus

$$213{\cdot}4 \div 21{\cdot}34 = \log 213{\cdot}4 - \log 21{\cdot}34$$
$$= 2{\cdot}3292 - 1{\cdot}3292 = 1{\cdot}00$$

Its antilog = 10·00 the required quotient.

By application of the rules for multiplication, we can also calculate Proportions by the log method. Here we add together the logarithms of the second and the third terms and from their sum subtract the logarithm of the first, and the remainder will be the logarithms of the fourth term.

To perform Involution by logarithms we multiply the logarithms of the given number by the exponent of the power to which it is to be raised and the product will be the logarithms of the required power. Thus the square of 21·34 will be calculated as follows :

$$(21{\cdot}34)^2 = \log (21{\cdot}34)^2 = 2 \times 1{\cdot}3292$$
$$= 2{\cdot}6584.$$

Its antilog = 455·4 the required square.

To perform Evolution by logarithms we divide the logarithms of the given number by the exponent of the root which is to be extracted, and the quotient will be the logarithms of the required root. Thus

$$\sqrt{21{\cdot}34} = \tfrac{1}{2} \log 21{\cdot}34 = \tfrac{1}{2} \times 1{\cdot}3292$$
$$= 0{\cdot}6646.$$

Its antilog = 4·619 the required square root.

Books invaluable for all statistical workers are :

1. *Mills—Statistical Methods.*
2. *Yule—Introduction to the Theory of Statistics.*
3. *Castle—Four Figure Mathematical Tables.*
4. *Pearson—Tables for Statisticians and Biometricians.*
5. *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals of all Integral Numbers up to* 10,000.

Powers, Roots and Reciprocals.

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	$\frac{1}{n}$	n	n_2	$\sqrt{n}$	$\sqrt{10n}$	$\frac{1}{n}$
1	1	1·000	3·162	1	51	2601	7·141	22·583	·0196
2	4	1·414	4·472	·5000	52	2704	7·211	22·804	·0192
3	9	1·732	5·477	·3333	53	2809	7·280	23·022	·0189
4	16	2·000	6·325	·2500	54	2916	7·348	23·238	·0185
5	25	2·236	7·071	·2000	55	3025	7·416	23·452	·0182
6	36	2·449	7·746	·1667	56	3136	7·483	23·664	·0179
7	49	2·646	8·367	·1429	57	3249	7·550	23·875	·0175
8	64	2·828	8·944	·1250	58	3364	7·616	24·083	·0172
9	81	3·000	9·487	·1111	59	3481	7·681	24·290	·0169
10	100	3·162	10·000	·1000	60	3600	7·746	24·495	·0167
11	121	3·317	10·488	·0909	61	3721	7·810	24·698	·0164
12	144	3·464	10·954	·0833	62	3844	7·874	24·900	·0161
13	169	3·606	11·402	·0769	63	3969	7·937	25·100	·0159
14	196	3·742	11·832	·0714	64	4096	8·000	25·298	·0156
15	225	3·873	12·247	·0667	65	4225	8·062	25·495	·0154
16	256	4·000	12·649	·0625	66	4356	8·124	25·690	·0152
17	289	4·123	13·038	·0588	67	4489	8·185	25·884	·0149
18	324	4·243	13·416	·0556	68	4624	8·246	26·077	·0147
19	361	4·359	13·784	·0526	69	4761	8·307	26·268	·0145
20	400	4·472	14·142	·0500	70	4900	8·367	26·458	·0143
21	441	4·583	14·491	·0476	71	5041	8·426	26·646	·0141
22	484	4·690	14·832	·0455	72	5184	8·485	26·833	·0139
23	529	4·796	15·166	·0435	73	5329	8·544	27·019	·0137
24	576	4·899	15·492	·0417	74	5476	8·602	27·203	·0135
25	625	5·000	15·811	·0400	75	5625	8·660	27·386	·0133
26	676	5·099	16·125	·0385	76	5776	8·718	27·568	·0132
27	729	5·196	16·432	·0371	77	5929	8·775	27·749	·0130
28	784	5·292	16·733	·0358	78	6084	8·832	27·928	·0128
29	841	5·385	17·029	·0345	79	6241	8·888	28·107	·0127
30	900	5·477	17·321	·0333	80	6400	8·944	28·284	·0125
31	961	5·568	17·607	·0323	81	6561	9·000	28·460	·0123
32	1024	5·657	17·889	·0312	82	6724	9·055	28·636	·0122
33	1089	5·745	18·166	·0303	83	6889	9·110	28·810	·0120
34	1156	5·831	18·439	·0294	84	7056	9·165	28·983	·0119
35	1225	5·916	18·708	·0286	85	7225	9·220	29·155	·0118
36	1296	6·000	18·974	·0278	86	7396	9·274	29·326	·0116
37	1369	6·083	19·235	·0270	87	7569	9·327	29·496	·0115
38	1444	6·164	19·494	·0263	88	7744	9·381	29·665	·0114
39	1521	6·245	19·748	·0256	89	7921	9·434	29·833	·0113
40	1600	6·325	20·000	·0250	90	8100	9·487	30·000	·0111
41	1681	6·403	20·248	·0244	91	8281	9·539	30·166	·0110
42	1764	6·481	20·494	·0238	92	8464	9·592	30·332	·0109
43	1849	6·557	20·736	·0233	93	8649	9·644	30·496	·0107
44	1936	6·633	20·976	·0227	94	8836	9·695	30·659	·0106
45	2025	6·708	21·213	·0222	95	9025	9·747	30·822	·0105
46	2116	6·782	21·448	·0217	96	9216	9·798	30·984	·0104
47	2209	6·856	21·679	·0213	97	9409	9·849	31·145	·0103
48	2304	6·928	21·909	·0208	98	9604	9·899	31·305	·0102
49	2401	7·000	22·136	·0204	99	9801	9·950	31·464	·0101
50	2500	7·071	22·361	·0200	100	10000	10·000	31·623	·0100

Conversion of r into z and z into r.

r	For z add	z	For r subtract
0·000-0·114	0·000	0·000-0·114	0·000
0·115-0·163	0·001	0·115-0·165	0·001
0·164-0·194	0·002	0·166-0·196	0·002
0·195-0·216	0·003	0·197-0·220	0·003
0·217-0·235	0·004	0·221-0·240	0·004
0·236-0·251	0·005	0·241-0·256	0·005
0·252-0·265	0·006	0·257-0·271	0·006
0·266-0·277	0·007	0·272-0·285	0·007
0·278-0·288	0·008	0·286-0·297	0·008
0·289-0·299	0·009	0·298-0·309	0·009
0·300-0·309	0·010	0·310-0·320	0·010
0·310-0·318	0·011	0·321-0·330	0·011
0·319-0·327	0·012	0·331-0·339	0·012
0·328-0·335	0·013	0·340-0·348	0·013
0·336-0·343	0·014	0·349-0·357	0·014
0·344-0·350	0·015	0·358-0·365	0·015
0·351-0·357	0·016	0·366-0·373	0·016
3·358-0·364	0·017	0·374-0·381	0·017
0·365-0·371	0·018	0·382-0·389	0·018
0·372-0·377	0·019	0·390-0·396	0·019
0·378-0·383	0·020	0·397-0·403	0·020
0·384-0·388	0·021	0·404-0·409	0·021
0·389-0·393	0·022	0·410-0·416	0·022
0·394-0·399	0·023	0·417-0·422	0·023
0·400-0·404	0·024	0·423-0·428	0·024
0·405-0·409	0·025	0·429-0·434	0·025
0·410-0·414	0·026	0·435-0·440	0·026
0·415-0·419	0·027	0·441-0·446	0·027
0·420-0·423	0·028	0·447-0·452	0·028
0·424-0·428	0·029	0·453-0·457	0·029
0·429-0·432	0·030	0·458-0·463	0·030
0·433-0·436	0·031	0·464-0·468	0·031
0·437-0·441	0·032	0·469-0·473	0·032
0·442-0·445	0·033	0·474-0·478	0·033
0·446-0·449	0·034	0·479-0·483	0·034
0·450-0·453	0·035	0·484-0·488	0·035
0·454-0·456	0·036	0·489-0·493	0·036
0·457-0·460	0·037	0·494-0·498	0·037
0·461-0·464	0·038	0·499-0·502	0·038
0·465-0·467	0·039	0·503-0·507	0·039
0·468-0·471	0·040	0·508-0·512	0·040
0·472-0·474	0·041	0·513-0·516	0·041
0·475-0·478	0·042	0·517-0·520	0·042
0·479-0·481	0·043	0·521-0·525	0·043
0·482-0·484	0·044	0·526-0·529	0·044
0·485-0·488	0·045	0·530-0·533	0·045
0·489-0·491	0·046	0·534-0·537	0·046
0·492-0·494	0·047	0·538-0·542	0·047
0·495-0·497	0·048	0·543-0·546	0·048
0·498-0·500	0·049	0·547-0·550	0·049

To use this table, look up the Value of r in the left-hand column and add to it the corresponding value in the second column, as z is always bigger than r. To turn z into r, look up z in the third column and subtract the corresponding entry in the last column.